
THE UNIVERSITY *of* EDINBURGH

# Synthesising Conversational Speech Using Found Data

*Johannah O'Mahony*

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2024

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Johannah O'Mahony*)

# Acknowledgments

First, I would like to thank my supervisor Simon King for all of his support over the last four years. I couldn't have asked for a better person to oversee my research! Thank you for providing guidance at every step of the way and for reminding me that while research is never finished, a PhD should finish! To Catherine Lai, my second supervisor – I knew after our first meeting that it would be amazing to work with you. I've enjoyed every chat we have had about the complexities of conversation and prosody, and I hope to have more in the future. Thank you for all of your academic support and guidance *and* for being the witness at my wedding! I would also like to thank my examiners Petra Wagner and Korin Richmond for taking the time to read my thesis, for providing detailed and invaluable feedback and for making the viva a really enjoyable experience!

To the phonetics research group at the University of Helsinki for making me feel at home during the course of my research stay! A particular shout-out to Juraj Šimko, Martti Vainio, Antti Suni, Sofoklis Kakouros, Heini Kallio! Thank you to Esther Klabbers for supervising my industry project. To my collaborators, Adriana Stan, Mikey Elmers, Sofoklis Kakouros and Éva Székely – it was amazing getting to work with all of you!

I would also like to thank all of the senior members of the Marie Skłodowska-Curie Innovative Training Network "Conversational Brains" (CoBra) for organising our training events and providing academic guidance throughout the last four years. Thank you to the CoBra PhD cohort for making every event interesting and enjoyable, in particular to Greta, Lena, Joanna, Tom, Carol and Adaeze!

To the students that I had the pleasure of supervising: Niamh, Xi, Wenjing and David. Working as a supervisor has truly been my favourite part of the last four years.

To everyone at CSTR for being the most lovely research group! To my amazing friends from my time in Edinburgh: Sarenne, Dan, Atli, Adaeze, Jacob, Jason, Irene – it was really hard to say goodbye to Edinburgh because of all of you! To my amazing friends abroad: Ghyslaine, Wiske, Femke, Maud, Miki, Shivara, Sofia, Aza, Ben, Brónagh, Siobhán and Erika - thank you for your support and for being a welcome distraction during the PhD. Thank you to my parents, Catherine and Derek, and my brother Adam for your support and for learning quickly that it is best not to keep asking how my thesis is going. Last but not least: Chau, thank you for everything!

# Lay Summary

Synthetic speech is increasingly found in technology that we use in our daily lives, for example voice assistants, like Siri or Alexa. But the speech in many applications still doesn't sound quite like a human would, for example, it can sound robotic or the intonation might not match the conversational context. Synthesising conversational speech, the form of speech that we use in our daily interactions, is challenging. First, when training speech synthesis models, we need high quality data, but such high-quality datasets are often not available for conversational speech. Second, conversational data is highly variable and complex. For example, when we talk to one another, we often hesitate, use fillers like *uhm* or *uh*, laugh as we speak, or talk over one another. Third, when we speak in a conversation, each utterance is embedded in a specific context, and this context affects what intonation or melodic properties we use when replying to someone. Finally, for many years, we have studied the intonation and melodic properties of speech using read speech utterances. This means that we still have much to learn about how speakers talk in natural conversation.

In this thesis, we address the issues above in three parts. In Part 1, we examine how listeners evaluate the quality of synthetic speech when it is embedded in a context. We find that the instructions given to listeners affect the ratings of synthetic speech. We also find that presenting synthetic speech in context increases the ratings of synthetic speech compared to synthetic speech presented in isolation. We conclude that the methods used for evaluating synthetic speech in isolation may not be suitable for rating speech in context.

In Part 2, we introduce two methods that can be used to both improve conversational speech synthesis and study conversational prosody. In the first method, we use a dataset of conversational podcasts and a dataset of read speech to train a speech synthesis model. We investigate whether the addition of real conversational speech data during training can help a model sound to more conversational. We find that training a model on a mixture of read speech and conversational speech enables the system to generate questions that are more preferred by listeners to questions synthesised by a model trained only on read speech data. In the second method, we create a controllable speech synthesis model using a linguistically-motivated polynomial representation of $F_0$ that can model intonational patterns used by speakers. Providing this representation as additional information to a speech synthesis model during training allows us to change the intonation of the synthetic speech at inference time, and in turn allows us to create different renditions of the same textual content for use in perception experiments.

In Part 3, we present two case studies which examine the impact of context on how we melodically realise an utterance in conversation. In the first study, we use found data and the polynomial representations from Part 2 to explore the different ways that the discourse marker "well" is realised by speakers. We use a clustering method to group similar renditions

of the discourse marker "well" in a large conversational dataset. We then take the average features of each group and use these features to control a speech synthesis model, creating 20 renditions of the same texts: *well yes* and *well no*. We present each of the 20 renditions of the above texts to listeners in a perception experiment to investigate how the melodic properties of the discourse marker affect the perceived attitude of a speaker. We find that the prosody of the discourse marker affects the level of agreement or disagreement perceived by listeners.

In the second study, we use found conversational data to investigate the melodic properties of turn-taking in conversation. We provide additional turn-taking information for each utterance during training. Specifically, we give information about whether each utterance was the final utterance in a speaker's turn, or whether the speaker continued to speak after the utterance. We find that providing this information to the model during training allows us to synthesise different renditions of the same text, one which sounds more turn-final and the other which sounds more turn-medial. This effect is found in three out of five synthetic voices tested. We also show that this method can be used to examine global prosodic tendencies in conversational data, allowing us to study the prosody of conversation using speech synthesis.

# Abstract

End-to-end speech synthesis models perform well when trained with clean read speech data. Modelling conversational speech, the form of speech that we use every day, however, is more challenging. First, we lack high-quality conversational datasets that are suitable for training speech synthesis models. Second, conversational data is highly variable, containing challenging spontaneous phenomena, such as overlapping speech and laughter. Third, each conversational utterance is embedded in a communicative context, and there are many contextual factors which must be accounted for. Finally, there exists a significant knowledge gap with respect to our understanding of both speech perception and speech production in context, both in the fields of speech technology and speech science.

In this work, we addressed these issues in three parts. In Part 1, we examined three factors that potentially affect the evaluation of speech synthesis output in context, namely the task instructions, between-sentence textual dependency and the prosodic realisation of the utterances. We found that task instructions can affect ratings, and we found that presenting speech in context narrows the gap of Mean Opinion Scores between the contextually appropriate utterance and the non-appropriate utterance. This suggests that MOS might not be sufficiently sensitive to evaluate speech synthesis in context. We conclude that more targeted evaluation is necessary to capture contextual effects.

In Part 2, we present two studies on improving conversational prosody using found data and controllable synthesis. In the first study, we find that training a model on a data mixture of found conversational speech (questions and answers) and read speech can improve the realisation of questions as measured by an increase in preferences for our datamix model over the baseline, which was only trained on read speech. For answers, no significant difference between the systems was found. In the second study, we used a linguistically-motivated word-level $F_0$ representations based on Legendre Polynomial coefficients to condition a FastPitch model, allowing us to control the intonation of an utterance. We found that conditioning a model on these representations increases to similarity of the $F_0$ contours between the system output and the target output over the baseline and a categorically-conditioned model. The proposed representations can then be used to explore patterns in conversational speech.

In Part 3, we present two case studies investigating the impact of context on an utterance's prosodic realisation. In the first study, we used found data and our intonation representations from Part 2 to explore prosodic variation on the discourse marker "well". Using clusters from the data exploration, we synthesised 20 different renditions of a positive polarity utterance, *well yes*, and a negative polarity utterance, *well no*, and performed a listening test to assess the degree of agreement perceived by listeners. We found that the prosodic rendition of the utterance can affect the perceived agreement or

disagreement of the speaker highlighting an example of the prosody-pragmatics interface. In the second study, we used found data to explore turn-taking cues in conversation. We found that conditioning a FastPitch speech synthesis model on turn-taking information leads to perceptible differences in the turn-finality of an utterance as measured in subjective listening tests. We showed that we can use speech synthesis to generate stimuli which reflect the global trends in the training data and that this method can complement corpus research in phonetics.

# Datasets

**CANDOR Corpus** Reece et al. (2023).

**LJ Speech dataset** Ito and Johnson (2017).

**Switchboard Corpus I** Godfrey and Holliman (1993).

**The Spotify 100 000 Podcast Dataset** Clifton et al. (2020).

# List of Figures

# List of Tables

# Acronyms

$F_0$ fundamental frequency

**ASR** Automatic Speech Recognition

**CA** Conversation Analysis

**CLMM** cumulative link mixed-model

**CNN** convolutional neural network

**CWT** Continuous Wavelet Transform

**DA** Dialogue Act

**DM** Discourse Marker

**E2E** end-to-end

**G2P** Grapheme-to-phoneme

**GMM** Gaussian mixture model

**GST** Global Style Tokens

**HMM** Hidden Markov Model

**HNR** noise-to-harmonics ratio

**IPU** inter-pausal unit

**LNRE** large number of rare events

**MFA** Montreal Forced Aligner

**MOS** Mean Opinion Score

**MUSHRA** Multiple Stimuli with Hidden Reference and Anchor

**NLP** Natural Language Processing

**PaIntE** Parametric representation of Intonation Events model

**POS** Part of Speech

**RMSE** Root Mean Squared Error

**RNN** recurrent neural network

**SNR** speech-to-noise ratio

**TD-PSOLA** Time-Domain Pitch-Synchronous Overlap-and-Add

**TTS** Text-to-Speech

**VAE** variational autoencoder

# Contents

# Part I

# Background

# 1

# Introduction

## 1.1 Motivation

Every day, speakers engage in the joint activity of conversation. In fact, conversation is the primary activity in which language is used (Schegloff, 1999; Clark, 1996a). It is therefore unsurprising that we have long endeavoured to create computational systems, such as spoken dialogue systems (SDS), and in particular *conversational dialogue systems* (Skantze, 2007), with which we can converse. As suggested by the name, a *spoken* dialogue system requires speech to be generated to facilitate interaction. This is achieved by using a Text-to-Speech (TTS) model, which after receiving the desired text to be spoken, produces a speech waveform. One of the potential goals when creating SDS systems, is to make them *natural* or more *human-like* (Edlund et al., 2008) and thus when synthesising speech for such a system, it too should meet those requirements.

Since the advent of end-to-end (E2E) speech synthesis models, the quality of synthetic speech has improved immensely, reaching high levels of *naturalness*, close to that of recordings of human speech (e.g. Tan et al. (2024)). However, these improvements have largely been seen when training on and synthesising recordings of isolated read speech utterances. Importantly, read speech does not reflect the naturally occurring speech found in conversation, which is spontaneous or unplanned, interactive and highly context-dependent (Campbell, 1997). Thus in the endeavour to create ever more *natural* speech, for use in spoken dialogue systems, we should ideally synthesise speech which resembles that which is found in real conversations (Székely et al., 2019b; Dall, 2017). This will mean moving from synthesising isolated utterances, to synthesising a felicitous utterance given its communicative intent in the wider communicative context.

As mentioned, conversational speech is spontaneous and found in interactive settings. It therefore exhibits multiple differences to the phonetically-balanced read speech that has traditionally been used to train speech synthesis models (Campbell, 2006). The spontaneous nature of conversational speech means that it contains phenomena related to speech planning, some of which have received attention in previous speech synthesis research. Examples include false starts or hesitations (Betz et al., 2018), filled pauses (Dall et al., 2016b; Székely et al., 2019a) or incomplete utterances. Conversational speech can also contain paralinguistic phenomena, such as laughter or smiling voice (Kirkland et al.,

2021). The interactional nature of conversation means that conversational speech contains phenomena which facilitate interaction, such as the use of particular dialogue acts, turn-taking cues, backchannels and specific discourse markers which help to structure the unfolding sequence of conversational turns. While all of these phenomena are aspects of natural conversational speech, their inclusion in an SDS system will likely be specific to the needs of the application.

Next to the aforementioned differences, and importantly for speech synthesis, conversational speech also exhibits differing phonetic characteristics to those of read speech. For example, conversational speech is a form of spontaneous speech and spontaneous speech contains large amounts phonetic reduction and variability (Tucker and Mukai, 2023). Therefore, the phonetic realisation of words in conversational speech may differ significantly from the canonical transcriptions (Ernestus, 2011) that are found in a standard lexicon used for training speech synthesis models with read speech. Further, both read and conversational speech can differ in their prosody (Hazan and Baker, 2010; Adigwe and Klabbers, 2022; Andersson, 2013). Here, prosody refers to suprasegmental aspects of the speech signal, such as rhythm, duration, pitch and intensity. The differences in the prosodic realisation between read and spontaneous speech may reflect the differing roles prosody plays in both speech modalities.

In this thesis, we are primarily concerned with synthesising appropriate conversational prosody and studying its role in facilitating conversational interaction and signalling pragmatic meaning. Prosody facilitates conversation by signalling turn-taking cues, the stance or attitude of a speaker, discourse relations and discourse meaning (Cole, 2015), among other functions. Importantly, all of these functions are expressed through the same speech signal and at the same time (House, 2006). Further, prosody serves as a contextualisation cue in conversation to aid the inferential processes that listeners use when comprehending an utterance in context (Gumperz, 1992; Couper-Kuhlen and Selting, 1996), and the prosodic realisation is both shaped by the communicative context, but also itself serves to shape the context of the conversation (Heritage, 1984). Thus, the current goal of synthesising conversational TTS is to produce speech that matches the desired communicative effect of an utterance in a conversation. This is a non-trivial task and will involve research into the prosody of naturally-occurring conversation.

There are a number of challenges to synthesising conversational speech, summarised here, but described in more detail in the following section. First, there are few high-quality conversational datasets which are appropriate for training conversational speech synthesis models. This has led to the emergence of work training TTS models on *found* conversational data, an approach we too will take in this thesis. Second, when working with found conversational data, we are faced with an immense amount of phonetic and prosodic variation, and this variation needs to be accounted for (where possible) to enable systematic contextual features to be learned by the speech synthesis model. Third, our understanding of naturally occurring context-dependent conversational speech in speech science is limited. Finally, given that we have trained a conversational speech synthesis model, we are faced with the challenge of evaluating the model, and as of yet, suitable paradigms for conversational speech have not been developed.

This thesis is therefore concerned with conversation from both a speech synthesis and speech science perspective, as we believe that for the complex problems posed by conversational speech, we need an interdisciplinary approach. From a speech science perspective, we need new tools and methods to explore conversational speech using *real*

conversational data, and we believe speech synthesis models have many advantages over traditional methods of stimuli creation which can aid this exploration. From a speech synthesis perspective, we need tools to identify and characterise patterns in the data, as well as methods for the evaluation of contextualised conversational speech, which ultimately speech scientists can facilitate.

## 1.2 Challenges to Synthesising Conversational Speech

As described above, synthesising conversational speech presents a number of challenges. In this section, we will explore four main challenges faced when attempting to synthesise conversational speech.

### 1.2.1 Data

The first challenge that we are faced with when attempting to synthesise conversational speech is a **data sparsity issue**. Traditionally, to synthesise high-quality speech, databases of read prompts were recorded by a voice talent in ideal recording conditions (Campbell, 2006). Thus, data sparsity refers not to the availability of data per se, but to the availability of high-quality data with which we can train a conversational TTS model. Most large-scale conversational speech corpora have been recorded for the purposes of training Automatic Speech Recognition (ASR) systems (e.g. Switchboard Corpus I (Godfrey and Holliman, 1993), CALLHOME (Canavan et al., 1997) and the Fisher Corpus (Cieri et al., 2004)). These corpora were recorded via the telephone, thus in sub-optimal recording conditions. Further, though these corpora contain a large number conversations, for example Switchboard Corpus I contains 2,400 conversations and the Fisher Corpus contains 5,850 conversations, both corpora have a limited quantity of data per speaker, as conversations are roughly 10 minutes in duration.

Because of the scarcity of high-quality conversational speech, work has started to synthesise speech using *found data* in the hope of finding recordings with sufficient recording quality and potentially larger quantities of data per speaker. In this thesis, *found* data refers to data which has not been recorded for the purposes of training speech synthesis models (Saeki et al., 2024). Examples of found data include podcasts (Szekely et al., 2019; Székely et al., 2019b), audiobooks (Zen et al., 2019), or audio from Youtube videos (Chen et al., 2023). The benefit of using found conversational data is that we can potentially acquire large quantities of speech data, which capture more speaking styles and conversational contexts. The disadvantage of this approach, is that we are faced with a boundless amount of variation in the data (Ogden, 2007).

### 1.2.2 Variation

Given that we have chosen to use found conversational data, we are then faced with the second challenge when synthesising conversational speech: **accounting for variation in the data**. This variation stems from the *context* in which the speech occurs, and many aspects of context exert influence on the phonetic form of the message. We can identify different aspects of context which will affect the realisation speech in the data proposed by Cruttenden (1997a) as cited by (Ogden, 2006, p1753) : the *sociolinguistic* context, the

*attitudinal* context, and the *discourse context*. Additionally, for the purposes of training speech synthesis models, we can also identify the *situational* context and the non-linguistic *recording* context. However these are merely high-level aspects of context, each of which will subsume many different linguistic and paralinguistic phenomena, which we will describe in the next chapter. Crucially, and as we will see, these aspects of context are not independent of one another – they are closely entangled, exerting influence on the same shared speech signal – and thus the more aspects of context we can account for, the more we can begin to understand their interaction.

### 1.2.3 Knowledge gap

In speech science, the primary form of speech which has been studied is carefully controlled lab-speech (Wagner et al., 2015b; Swerts and Hirschberg, 1998). For example Wagner et al. (2015b) found that in 2011, 68% of work presented at ICPhS[1] used scripted speech, and more recently Cangemi et al. (2023) found that at Speech Prosody 2020 90% of work on the prosody of questions involved lab-elicited speech (Cangemi et al., 2023). This limitation has been repeatedly criticised, especially in the context of Conversation Analysis (CA), and in particular, by practitioners of *interactional prosody* (e.g. Couper-Kuhlen and Selting (1996); Swerts and Hirschberg (1998); Culpeper (2011)). What this means is that we still lack fundamental understanding of the effect of the interactional context on the prosodic realisation an utterance. Because of this, theories of speech perception and production, which have been developed based on lab-elicited speech, might not generalise to real conversational interaction (Cangemi and Baumann, 2020; Swerts and Hirschberg, 1998). Thus we still have little understanding of the degree to which the various functions of prosody are systematic, and how well prosodic cues found in corpus studies impact listener perception.

In speech science, one of the difficulties when studying conversational speech, is that creating *controlled*, but *spontaneous* stimuli for perception experiments is difficult, if not impossible by definition. Thus most work in speech perception research has used *lab speech* based on scripted dialogue e.g. Cutler and Pearson (1985). Others manipulate existing recordings of conversational speech, although this approach might also lead to unnatural sounding speech that is not ecologically valid.

For speech synthesis, our lack of knowledge about prosody in interaction has knock-on effects for how we approach modelling interactive prosodic phenomena. Firstly, we do not know enough about the interplay between prosody, pragmatics and the interactional context, which makes modelling these phenomena challenging. For example, many aspects of meaning which are expressed in conversation, such as attitude or stance do not have clear taxonomies which can be used to label data, even if stance and attitude are more prevalent in conversation than prototypical emotions, such as anger and happiness (Campbell, 2006). As Ogden (2006) notes, this is because such subtle effects are highly subjective. Secondly, we do not know how speech perception changes as a function of context or what aspects of prosody a conversational system should learn. The perception and salience of prosodic cues is as important in speech synthesis research as the modelling of prosody itself because it impacts how we evaluate our models. This lack of understanding makes it difficult to address evaluating such systems.

[1] International Congress of Phonetic Sciences

### 1.2.4 Evaluation

A final challenge when synthesising conversational speech is the **evaluation of conversational speech synthesis**. Traditional evaluation paradigms, such as the Mean Opinion Score (MOS) test, were developed for isolated utterances. Furthermore, common metrics used in MOS, such as *naturalness* or *intelligibility*, may not be fit for purpose when evaluating differing prosodic renditions of an utterance in context. How do we construct appropriate evaluation paradigms, and how to we construct evaluation stimuli which consist of aspects of conversation which we would like to capture? Though this will likely depend on the use-case for the TTS voice, it is imperative that we define aspects of context and create appropriate evaluation for these context effects. This may mean identifying particular spoken material in corpora, and creating more targeted evaluation, or situated evaluation (Wagner et al., 2019).

### 1.2.5 Ethical Considerations

Next to the practical concerns above, when synthesising conversational speech using found data there are a number of ethical considerations that must be taken into account. The first regards the ethics of using of found data to train speech synthesis models. Though the name *found data* suggests that this data can be taken from any source, in this thesis found data means that this data was not recorded for the purposes of training speech synthesis models, that is to say, it is not recorded in optimal studio conditions and does not comprise isolated read out prompts of either written material or conversational material. In this thesis, the found data used comes from corpora that have been published for use in academic research and we do not use sources of conversational speech beyond these corpora.

Moreover, throughout the thesis, we have developed methods in which found data can be used to learn conversational patterns, without necessarily synthesising the individual speakers seen during training. For example, in Chapter 4, we use a datamixing strategy, mixing read speech and found data, to learn conversational patterns without specifically synthesising the speakers from the found dataset. In Chapter 6, we use found data to learn prosodic patterns used when using the discourse marker *well*, but again we do not synthesise individual speakers from the found data. An exception to this is found in Chapter 7. Here we do synthesise five speakers from the CANDOR corpus (Reece et al., 2023).

The final ethical issue concerns the creation of hyper-realistic voices for use in speech applications and spoken dialogue systems. While we often assert that we want to create ever more natural and human-like speech, we must acknowledge the dangers of creating synthetic speech which cannot be easily distinguished for real human speech. For example, synthetic speech can be maliciously used to impersonate others. In this thesis, we predominantly approach the use of found data in speech synthesis as a tool to enrich our understanding of the prosodic-pragmatic interface in the linguistic sciences in an academic setting. We will not release model checkpoints to the general public and do not condone the use of speech synthesis models to impersonate individuals or to otherwise cause harm. Next to the goal to enrich our understanding with speech science, we do believe that this knowledge can be used to enhance human-computer interaction within SDS systems, for example being able to manage conversational interaction through the use of prosodic cues, or enabling a system to ask questions with an appropriate prosodic realisation. This does

not entail that the resulting speech must be hyper-realistic or be able to synthesise all aspects of human behaviour in conversation.

## 1.3 Main Thesis Contributions

In the following section, we summarise the main contributions in this thesis.

- Chapter 3
  - We address the evaluation of synthetic speech in context by presenting three controlled experiments which we investigate 1) the role of instructions and 2) the context-dependency of utterances and 3) whether MOS is a suitable measure for rating speech in context.
  - We show that the MOS paradigm may not be suitable for rating speech in context.
- Chapter 4
  - We present a method for improving the realisation of under-represented dialogue acts in TTS by using a datamixing strategy in which we combine conversational *found data* with read speech data to train a model.
- Chapter 5
  - We develop a method for controlling the intonation of a speech synthesis model with word- and phrase-level parametric representations of $F_0$.
  - We show that this method can lead to increased similarity to a reference recording as judged by listeners, and as measured using objective evaluation.
- Chapter 6
  - We present a method for exploring conversational prosodic patterns in found data using the representations from Chapter 5.
  - We synthesise different prosodic renditions of utterances using the patterns found during exploration.
  - We show that this method can be used in speech science to create listening test stimuli in a data-driven way.
  - This shows that this method of stimuli creation can facilitate investigating the perception of conversational speech.
- Chapter 7
  - We investigate the prosody of turn-taking by conditioning a speech synthesis model on the position of an utterance in a turn.
  - We show that adding turn-position conditioning leads to an increase in turn-final judgements in certain speakers.
  - We show that this method can be complementary to traditional corpus studies and that it can both learn global characteristics of the training data, while also allowing us to synthesise the global characteristics of the data.

## 1.4 Summary of Chapters

### A: Background

#### Chapter 2: Background

In Chapter 2, we give an overview of conversation and the importance of communicative context in conversation. Further, we present the various functions of prosody in conversation. We then give an overview of the speech synthesis models that will be used in this thesis, and introduce concepts, such as controllability and evaluation in speech synthesis. Note that the background in this thesis is intended to *briefly* introduce some of the main concepts that will be relevant for this thesis, however each of the remaining chapters will contain a self-contained background section which is relevant to the particular study of that specific chapter.

### B: Evaluation in Context

#### Chapter 3: Evaluation of Speech Synthesis Output in Context

This chapter focuses evaluation of speech in context. Here we build on prior work by Clark et al. (2019) into speech synthesis evaluation in context using the MOS paradigm. In particular, we investigate one of the most interesting findings from the aforementioned study: utterances of synthetic speech were rated more highly in context than in isolation, even though they were not synthesised by a context-aware model. In this chapter, we focus on the perception of synthetic read speech in context. Specifically, we investigate three factors which might lead to increased ratings in context: the instructions, the textual content of the synthesised utterances and the prosodic realisation of the synthesised utterances.

### C: Speech Synthesis Methods

#### Chapter 4: Using Found Data to Improve Speech Synthesis Prosody

In this chapter, we highlight the difficulty in finding suitable data for training conversational speech synthesis models. We give an overview of previous data sources which have been used to train conversational speech synthesis models. We present a datamixing approach to improve the prosodic coverage of a read speech target speaker by enriching an existing corpus of read speech with conversational speech from found podcast data. Specifically, we enrich the dataset with questions and answers and evaluate whether a datamixing strategy can lead to increased preference in subjective listening tests.

#### Chapter 5: Hierarchical Intonation Control Using Legendre Polynomials

In Chapter 5, we explore a method of synthesising speech by conditioning a FastPitch model (Łańcucki, 2021) hierarchically on a sparse representation of the word-level $F_0$ contour and the phrasal-level $F_0$ slope, using Legendre Polynomials coefficients to approximate the $F_0$ contour. We motivate the use of this representation of $F_0$ by its use in previous linguistic studies (Grabe et al., 2007), by its low dimensionality and by the interpretability that the Legendre coefficients provide. We show that conditioning a model

on this sparse representation leads to prosodic renditions judged to be closer to a reference sample than a model which is conditioned on binary prominence and boundary labels and closer than the baseline model output.

### D: Case Studies

**Chapter 6: Exploring Pragmatic Variation of Discourse Markers Using Found Data**

This chapter focuses on discourse markers, which are linguistic elements that perform various functions in contextualising utterances within a conversation. In previous research, the functions of discourse markers have been found to be linked to their prosodic realisation. In this chapter, we explore the prosodic realisation of *well* using an unlabelled corpus of found conversational speech. Here, we use clustering to explore the variation in the prosodic realisation of *well* and we identify common prosodic patterns in a data-driven manner. The cluster centroids are synthesised using the controllable speech synthesis from Chapter 5 and are evaluated in relation the level of agreement/disagreement perceived by listeners. We therefore show how controllable speech synthesis can be used to create stimuli for perceptual experiments using patterns automatically found from real conversation.

**Chapter 7: Modelling Turn-taking Prosody from Found Data**

In this chapter, we use found conversational data to train a speech synthesis model that is conditioned on the turn-position of an utterance in a conversation to model turn-taking cues. The goal of this chapter is to use speech synthesis to explore whether speakers use prosodic means to signal turn transitions and whether these signals help listeners to judge whether a speaker will continue speaking or give up their turn. At the same time, in this chapter we explore the importance of data selection and the development of targeted evaluation sets for specific aspects of conversational speech that are being modelled.

## 1.5 List of Publications

### Primary Contributions

1. **O'Mahony, J.**, Oplustil Gallegos, P., Lai, C., & King, S. (2021). Factors Affecting the Evaluation of Synthetic Speech in Context. In *Proceedings of the 11th ISCA Speech Synthesis Workshop (SSW 11)* 148-153. doi:10.21437/SSW.2021-26

2. **O'Mahony, J.**, Lai, C., & King, S. (2022). Combining conversational speech with read speech to improve prosody in Text-to-Speech synthesis. In *Proceedings of Interspeech 2022* 3388-3392 doi:10.21437/Interspeech.2022-10167

3. **O'Mahony, J.**, Corkey, N. Lai, C., Klabbers, E., King, S. (2024) Hierarchical Intonation Modelling for Speech Synthesis using Legendre Polynomial Coefficients. In *Proceedings of Speech Prosody 2024*, 1030-1034, doi:10.21437/SpeechProsody.2024-208

4. **O'Mahony, J.**, Lai, C., & Székely, É. (2024). "Well", what can you do with messy data? Exploring the prosody and pragmatic function of the discourse marker "well" with found data and speech synthesis, *Proceedings of Interspeech 2024* (pp 4084-4088) doi:10.21437/Interspeech.2024-2122

5. **O'Mahony, J.**, Lai, C., King, S. (2023) Synthesising turn-taking cues using natural conversational data. *Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW12)*, 75-80, doi:10.21437/SSW.2023-12

## Secondary Contributions

1. Oplustil Gallegos, P., **O'Mahony, J.**, & King, S. (2021). Comparing acoustic and textual representations of previous linguistic context for improving Text-to-Speech. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)* (pp. 205-210). doi:10.21437/SSW.2021-36

2. Stan, A., & **O'Mahony, J.** (2023) An analysis on the effects of speaker embedding choice in non auto-regressive TTS. In *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*, 134-138, doi:10.21437/SSW.2023-21

3. Kakouros, S., & **O'Mahony, J.** (2023). What does BERT learn about prosody? In R. Skarnitzl, & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1454-1458). https://guarant.cz/icphs2023/622.pdf

4. Kruyt, J.*, Huttner, L.*, & **O'Mahony, J.** (2023)* *equal contribution. Investigating the relationship between prosodic entrainment and interaction style. In R. Skarnitzl, & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 3492-3496). https://guarant.cz/icphs2023/514.pdf

5. Elmers, M., **O'Mahony, J.**, & Székely, É. (2023). Synthesis after a couple PINTs: Investigating the role of pause-internal phonetic particles in speech synthesis and perception. *Proceedings of Interspeech 2023* 4843-4847. doi:10.21437/Interspeech.2023-2178

# 2

# Background

In this thesis, we approach modelling conversational speech synthesis with found data from two perspectives: 1) from a speech synthesis perspective and 2) from a speech science perspective. More specifically, from a speech synthesis perspective, we intend to use methods to illustrate certain aspects of conversation which should be considered when modelling conversational speech, and in doing so we conceptualise some of the difficulties with modelling conversational speech using found data. From the speech science perspective, we have chosen to use speech synthesis models as a tool which can aid the study of conversational prosody. In particular, we focus on methods for creating listening test stimuli with patterns which have been identified in real data, which can subsequently be synthesised using TTS.

The interdisciplinary approach that we take in this thesis is based on the fact that conversation is a difficult phenomenon to model, and also inspired by the work of Malisz et al. (2019) who call for more collaboration between disciplines. In particular, the fields speech science and speech synthesis have historically focused their attention on isolated read utterances. Thus, we still have limited understanding of the role of prosody in interactive contexts, and the degree to which this role is systematic. Therefore, to understand conversational speech, to synthesise conversational speech and importantly to evaluate whether we have succeeded, going forward, a collaborative approach between both disciplines is needed.

In this chapter, we will introduce the key concepts which will feature throughout this thesis. In Section 2.1, we give a brief overview of the structure of conversation and the notion of *context* in conversation. In Section 2.2, we introduce prosody, and aspects of conversation which have been found to affect how an utterance is prosodically realised. In Section 2.3, we introduce speech synthesis and the speech synthesis models which will be used in this thesis and in Section 2.3.4, we give an overview of approaches to conversational speech synthesis. Finally in Section 2.4, we give a brief overview of paradigms used to evaluate synthetic speech. Note that in this chapter, we will not give an exhaustive overview of each topic. In the chapters which follow, a detailed background will be given each of the topics explored.

## 2.1 Conversation

Conversation is spontaneous spoken interaction between two or more people (Clark, 1996a). But what precisely constitutes conversation is difficult to define (Schiffrin, 1990), as there are many settings in which humans interact through speech. Within CA, conversation can be subsumed under the category *talk-in-interaction* (Schegloff, 1999) or *speech exchange system* (Sacks et al., 1974). Other examples of *talk-in-interaction* include task-based dialogue, meetings, discussions and interviews (Schegloff, 1980), which can be distinguished from conversation based on various factors as outlined by Schegloff (1999), such as in their turn organisation, topic organisation, and sequence organisation. It is generally noted that conversation permits a less restrictive structure than that of a debate or interview (Heritage, 1984). Clark (1996b) states that conversation is found in personal settings:

> The spoken setting mentioned most often is conversation — either face-to-face or on the telephone. Conversations may be devoted to gossip, business transactions, or scientific matters, but they are all characterized by the free exchange of turns among the two or more participants. I will call these personal settings.
>
> (Clark, 1996b, p.4)

Like Schegloff (1999), Clark (1996b) later contrasts the above to *institutional* settings, such as meetings or proceedings in court, though he notes that these may also show overlapping characteristics with conversation, albeit with more constraints (Clark, 1996b, p.5).

Conversely, in the context of TTS research, conversation has not been strictly defined. The datasets used to synthesise "conversational" speech, can include interactions designed for voice agents for activities, such as as flight booking (Guo et al., 2021), interviews and task-oriented dialogue (Koriyama et al., 2010). Thus, it appears that conversation in speech synthesis research can be used as a term which can cover many different spoken interaction settings, and is therefore roughly equivalent to anything which is subsumed under talk-in-interaction. This means that many datasets used in speech synthesis for the purposes of synthesising conversational speech may not actually constitute *conversation*. However, as noted by Schegloff (1980), conversation might not necessarily be the best form of interaction to use depending on the intended application, for example a voice for a domain-specific dialogue system for flight booking may not need to be modelled using casual conversational data.

In this thesis, we take a more permissive approach to defining conversation than that of Clark (1996b) and Schegloff (1999). This is primarily due to the fact that when using found data, we do not always know the context in which the speech occurs. In this thesis, we define conversation as spontaneous spoken interaction between two or more people. Furthermore, in this thesis, we exclusively refer to two-party conversation rather than multi-party conversation where more than two participants take part. Therefore here, we have chosen to permit more instances of talk-in-interaction in my definition of conversation, such as podcast interviews, which will feature in Chapter 3. However, in the case studies presented in Chapter 6 and Chapter 7, we have focused on corpora of domain

unspecific *casual conversation*, akin to the definitions of Clark (1996b) and Schegloff (1999).

### 2.1.1 The Structure of Conversation

Conversation is a joint action in which participants collaborate to align their "representations of the situation under discussion" (Garrod and Pickering, 2009, p.295). Conversation exhibits a hierarchical structure which emerges over the course of the conversation (Clark, 1996a) as participants update their mental representations utterance by utterance. Though there is considerable variation in the literature with regards to the precise hierarchical units of conversation, there is consensus that conversation unfolds utterance-by-utterance in a sequence, with each utterance being relevant to what was previously said (Wilson and Wharton, 2006). An utterance's meaning is therefore inferred within the context it occurs (Heritage, 1984; Couper-Kuhlen and Selting, 1996).

As we have already alluded to, a central feature of conversation is that participants take turns to speak, and they do so in a coordinated fashion (Sacks et al., 1974). How speakers coordinate turn-taking is still an active area of research, but a number of pragmatic, syntactic, semantic and prosodic *turn-taking cues* have been identified (Skantze, 2021). That speakers take turns does not entail, however, that there are no occurrences of overlapping speech where both parties simultaneously produce speech. In fact, overlapping speech is frequent in conversation and can have different functions, such as providing short feedback utterances such as *yeah* and *uh-huh*, among others (see Schegloff (2000)).

Secondly, each turn in a conversation can be made up of several utterances or *chunks* (Szczepek Reed, 2010) which constitute the "'building blocks' of discourse structure" (Degand and Simon, 2009, p.2). There is no consensus with regards to the units of conversation or discourse (Szczepek Reed, 2010; Degand and Simon, 2009). There are therefore many ways to segment a conversation into smaller functional units some of which have theoretical motivations in discourse analysis or conversation analysis, and others which do not consider the linguistic or pragmatic content. Theoretical accounts have postulated discourse units based on intonation, semantic, syntactic and pragmatic considerations, but other approaches consider multi-dimensional units (Degand and Simon, 2009) which consider a combination of syntactic, prosodic and pragmatic features (Hu and Degand, 2023).

In quantitative corpus research, approaches to segmentation are often chosen which are not based on the conversational content of the utterances, but rather on features which can be automatically derived from the signal, such as silences (Koiso et al., 1998) or breaths (Szekely et al., 2019). The advantage of using such an operationalisation is that they are theory-neutral, require no human annotation and can be computed automatically (Koiso et al., 1998). These approaches therefore facilitate replicability across corpus studies.

In this thesis, we use the inter-pausal unit (IPU) to segment utterances in conversational speech[1] (see Figure. 2.1) in the absence of a model which can automatically detect other more meaningful units. An IPU is a stretch of speech which is delimited by silence on either side. A threshold of minimum silence between IPUs is chosen and in studies this threshold ranges between 50 - 500 ms, with most setting a threshold between

[1]Note that we do not use IPUs in Chapter 3 because the data used there is single channel and thus does not permit such segmentation.

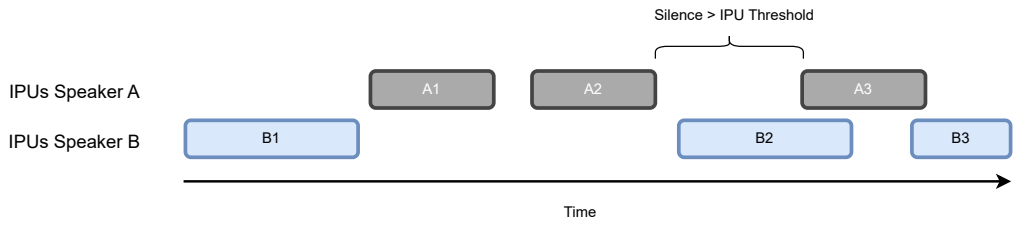


Figure 2.1: *Depiction of inter-pausal units (IPUs) in a conversation*

100 - 200 ms (Włodarczak and Wagner, 2013, p.1435). Importantly, as found in the study of Włodarczak and Wagner (2013), the threshold used when extracting IPUs can have consequences for results obtained in corpus studies, especially those quantifying conversational dynamics, such as quantifying duration of utterances and frequency of overlapping speech. We should also note that pauses have been found *within* units found in other segmentation approaches (Szczepek Reed, 2010), therefore the IPU should not necessarily constitute a theoretically-relevant minimal unit of conversation. We do not postulate that the IPU is necessarily the best way to segment conversation, but it mirrors previous approaches which have been used in corpus studies, for example into turn-taking (e.g. Brusco et al. (2017); Gravano et al. (2011)).

Finally, above the level of the utterance and turn, conversation exhibits larger hierarchical structures, such as topics, sections or joint projects (Clark, 1996a). For the sake of brevity, and because these notions do not feature heavily in this thesis, these larger units of conversation will be left out of scope.

### 2.1.2 Context in Conversation

As we saw, conversation unfolds in a sequential manner, utterance by utterance. A central concept in the study conversation within CA is *context* – what a speaker says and how they say it is dependent on the conversational context in which it occurs. The interpretation of an utterance is therefore also inferred from context (Heritage, 1984; Gumperz, 1992; House, 2007). Heritage (1984) as cited by Goodwin and Duranti (1992) states that each utterance in a conversation is "doubly contextual", that is to say, when an utterance is contributed by one of the speakers, it both is shaped by the previous conversational context, but itself also shapes or adds to the context (Heritage, 1984, p.242). Crucially, the prosody of an utterance has been found to function as a contextualising cue (Gumperz, 1992; House, 2007). As we will see in more detail in Section 2.2, prosody functions to contextualise utterances in many ways (Gumperz, 1992), including by signalling conversational structure (House, 2007), highlighting salient aspects of information (Gumperz, 1992), expressing paralinguistic information (House, 2007; Wilson and Wharton, 2006), such as stance or intent, among others.

Context has been studied from many different perspectives and there is no formal definition of context (Goodwin and Duranti, 1992). Therefore, here we situate the notion of context within the task at hand in this thesis, namely synthesising conversational speech from found data. As we have mentioned, the traditional paradigm in TTS is to synthesise *isolated* utterances, thus treating utterances as context-free entities. Goodwin and Duranti (1992) states that this has also been the primary approach in linguistics where an utterance

is treated "as though it had no ties to the talk that surrounds it" (Goodwin and Duranti, 1992, p.12). Goodwin and Duranti (1992) writes that treating utterances in isolation has been largely successful within linguistic sciences because it limited the object of study to something that could be clearly defined, for example studying syntactic structure within a sentence. The same can be said for the study of prosody in speech science, which also frequently operated on the utterance-level and was often constrained to speech which had been read out or elicited artificially (Couper-Kuhlen and Selting, 1996; Swerts and Hirschberg, 1998; Cangemi et al., 2023).

Treating utterances as isolated entities in TTS has also been largely unproblematic because traditionally TTS datasets comprised of recordings of isolated read out utterances. The utterances to be synthesised were therefore void of a communicative context, thus it was not necessary to take additional context into account. An example of this is the Arctic prompts which were a set of phonetically-balanced isolated utterances which were specifically designed for TTS (Kominek and Black, 2004). Conversely, when using found conversational speech, each utterance is embedded in a wider conversational context and thus the prosodic realisation will be affected by the conversational context in which it occurs and the realisation of the utterance will also serve to shape the upcoming context. If we were to segment a conversation into individual utterances and treat them as isolated entities it would limit the prosodic variation that we could potentially synthesise. The TTS model would learn the average characteristics of utterances in the dataset (Hodari, 2022), which would not take advantage of the rich amount of information they convey about conversational dynamics which are realised in the prosodic signal.

Recently, due to advances in TTS models, we can begin to consider conditioning models on additional context. This is evidenced by the growing number of papers which are beginning to use the previous conversational context to condition the synthesis of utterances (e.g Guo et al. (2021); Cong et al. (2021)). This is not to say, however, that context was never considered in early TTS research:

> Enormous advances in artificial intelligence will be needed before we can expect a computer to mimic people in using a variety of intonation patterns to express a variety of relations between the current utterance and the sense of the previous discourse.
>
> (Pierrehumbert, 1981, p.985)

Given that the prosodic form of an utterance both shapes and is shaped by the conversational context, a central question when synthesising speech using conversational found data is therefore what aspects of context to include in a model and how to incorporate context into models of conversational speech synthesis. In the next section, we therefore give an overview of aspects of prosody which are shaped by context and which serve too to shape the new conversational context.

## 2.2 Prosody in Conversation

### 2.2.1 Prosody

Prosody describes the aspects of speech which operate above the segmental level. Here we am referring to the aspects of speech which encapsulate pitch, duration and loudness

(Cruttenden, 1997b). The features just mentioned are perceptual constructs (Cole, 2015; Cruttenden, 1997b) but have correlates which can be measured in the speech signal: fundamental frequency ($F_0$), duration of phones, syllables and words, and intensity respectively. Next to these three features, though not heavily featured in this thesis, prosody also includes aspects of speech which affect the perception of voice quality, such as creaky voice or modal voice, which too can be measured in various ways including using jitter, shimmer and noise-to-harmonics ratio (HNR).

Before we explore the function of prosody in conversation, it is helpful to introduce aspects of prosodic structure which will be featured in this thesis. In this work, we mainly make reference to three aspects of prosodic structure: prosodic prominence, prosodic boundaries and prosodic phrases. As mentioned by Wagner et al. (2015a) for the concept prosodic prominence, and similarly by Schuetze-Coburn (1992) for the notion of the prosodic phrase, there are many approaches to defining these concepts which are grounded in different scientific practices. Therefore, we will define these terms with regards to their use in this thesis.

First, we refer to word-level prosodic prominence as a word acoustically standing out relative to other words, as measured by $F_0$, duration and intensity. Specifically, to estimate prosodic prominence, we use the Prosody Wavelet Toolkit[2] (Suni et al., 2017), which combines signals of duration, intensity and $F_0$ over which the Continuous Wavelet Transform (CWT) is calculated. The CWT is used to decompose the combined signal into scales of differing resolutions which capture the hierarchical structure of prosody i.e. from syllable to large units such as phrases (Suni et al., 2017). The toolkit can be applied to any language for which paired audio and word and phone transcriptions are present, though the importance of the prosodic parameters used when estimating prominence may need to be tuned for specific languages and speakers. To estimate prominence, local maxima across the scalogram are used to identify prominence (Suni et al., 2017). In some parts of the thesis, prominence will be treated as a binary distinction, prominent or not prominent, while in our proposed approach in Chapter 4, we model the $F_0$ shape on prominence words. Next to prosodic prominence, we will refer to prosodic phrases, which in this thesis is a prosodic unit delimited by the presence of a prosodic boundary. Prosodic boundaries, in this thesis, are approached in much the same way as prominence using the toolkit described above. Using the same decomposition of the combined signals of $F_0$, intensity and duration, local *minima* are used to detect boundaries (Suni et al., 2017).

### 2.2.2 Prosody in Context and Interaction

Prosody has many functions in conversation and there exists a wide number of taxonomies to describe the functions of prosody in conversation. In her comprehensive review of prosody in context, Cole (2015), for example, describes three main functions of prosody, namely signalling structure, signalling discourse meaning, and signalling information about the situational context. For example, under structure, Cole (2015) states that prosody signals the boundaries of words in an utterance and that prosody has a probabilistic relationship with signalling syntactic boundaries and signalling the beginnings and ends of discourse units. According to Cole (2015), discourse meaning can be encoded through prosody through the prominence of words in the discourse context

[2] `https://github.com/asuni/wavelet_prosody_toolkit`

signalling information structure, through signalling dialogue acts or speech acts, by signalling affective meaning, such as emotion or stance, and in managing interaction through the use of backchannels and signalling turn-taking cues. Finally Cole (2015), describes how prosody can shape information about the situational context by signalling information about the speaker, such as their dialect or their own idiosyncratic way of speaking. Further, prosody will change as a function of the situation in which speakers find themselves in, for example a formal discussion versus a casual chat.

The overview of Cole (2015), provides an excellent reference point for the various functions of prosody which we may wish to consider when modelling conversational speech. Crucially, these functions do not have a one-to-one correspondence with prosodic features in conversation. All of these functions will be expressed though a shared prosodic signal, and will show considerable interaction, as well as an interaction with the sequential context in which they occur (House, 2006). Furthermore, and as noted by Couper-Kuhlen and Selting (1996); Couper-Kuhlen (2005); Cruttenden (1984), identifying particular tunes which signal particular functions may not be a fruitful line of inquiry because the same intonational tune can mean very different things in different conversational contexts. For example, if in TTS research we were to segment a conversation and apply labels of meaning to every individual utterance, we might find, when conditioning a model on these labels, that no systematic patterns are found. This is because similar prosodic features can signal different meanings depending on the context in which they are found (Cruttenden, 1984). Therefore when modelling conversational speech, we ideally need to include context beyond the current utterance.

Prosody in context has been studied extensively in the field of CA. Within CA, prosody functions as a contextualisation cue (Gumperz, 1992; Couper-Kuhlen, 2005; Couper-Kuhlen and Selting, 1996). As we will see in Section 2.3.4, within TTS research, work on contextualised conversational speech has focused solely on conditioning speech synthesis models on the context of the previous utterances in the conversation. Current work on contextualised TTS does not focus on modelling how the speaker wishes to shape the context through their use of prosody as a contextualisation cue. As noted by Campbell (2006), "We recognize that an utterance has a direct relationship to a discourse event, and that the relationship is subject to constraints from two dimensions related to the factors; i.e., 1) influences from the speaker's own states (Self), and 2) influences from the listener and the discourse context (Other)" (Campbell, 2006, p.1174). Therefore, to model the dimension of *Self*, we need to identify speaker intent, and this is a non-trivial task. This also implies that when synthesising conversational TTS utterances, not only should an utterance be appropriate in the context as is often stated, but it should also appropriately manipulate the new conversational context in a way that the speaker intended.

In this thesis, we therefore focus on specific properties of utterances which speakers use to contextualise their own utterance. In Chapter 6, we investigate the role of prosody on discourse markers. We choose discourse markers because one of their main functions is to contextualise the target utterance within the context that has already unfolded (Brinton, 1996). Further, discourse markers, such as *well*, can convey paralinguistic information about the speaker's stance. They therefore provide an important cue for assessing what Campbell (2006) calls the dimension of Self. Secondly, in Chapter 7, we investigate turn-taking cues. Again, these are cues used by the speaker of the target utterance to facilitate the structure of conversation and again are a window into the Self dimension, as they are chosen by the speaker to indicate whether they will continue speaking or will give

up the floor.

## 2.3 Speech Synthesis

### 2.3.1 Overview of Speech Synthesis

In its simplest form, TTS entails the conversion of an input, usually a string of text, into a speech waveform. One of the biggest challenges in TTS is that the text to be synthesised is *under-specified* (Pierrehumbert, 1981). This underspecification is due to the fact that the textual content does not specify *how* something should be said – there are potentially infinite ways to acoustically realise a single text. For example, as we have seen, conversational speech is highly context-dependent and therefore speakers realise utterances in a wide variety of ways depending on a multitude of contextual factors, some of which were discussed in Section 2.2. However, most TTS systems synthesise isolated utterances. Therefore, in most approaches to TTS, determining how a text should be acoustically realised is solely based on a linguistic specification that is derived from the text in isolation.

In many TTS systems, the linguistic specification is derived from the text using a linguistic front-end. The goal of the linguistic front-end is to provide a richer representation of the text which can in turn be mapped to acoustic features using the back-end acoustic model. Typically, the front-end consists of many modules, each of which carries out different levels of linguistic analysis on the text. Common tasks in the front-end include: *text normalisation*, for example converting a string such as "$2.50" into a spoken form "*two dollars fifty*"; Grapheme-to-phoneme (G2P) conversion, that is converting the textual sequence into phones; and Part of Speech (POS) tagging. Additionally, for prosody, additional models predict the presence of prosodic features, for example the placement of prosodic prominence and the delimitation of prosodic phrases (Taylor, 2009) to provide a richer specification of how the text should be spoken.

Once a linguistic specification has been derived from the input text, the back-end is responsible for mapping this specification to acoustic features using an acoustic model. There have been several approaches to acoustic modelling in TTS, including statistical parametric approaches like Hidden Markov Model (HMM) synthesis and neural network-based approaches (Wu et al., 2016). Additionally, some acoustic back-ends will require more than one model to synthesise speech, for example an HMM-based synthesis model usually requires a further duration model.

More recently, some researchers have begun to use neural sequence-to-sequence models in speech synthesis research (e.g. Wang et al. (2017); Ren et al. (2019, 2021); Łańcucki (2021)). Modelling TTS as a sequence-to-sequence task entails mapping an input, for example text, directly to an acoustic representation, for example a mel spectrogram. This is usually achieved using an attention-based encoder-decoder architecture, in which the encoder creates a rich sequential representation of the input string to be synthesised and the decoder predicts the acoustic features from the output hidden representation of the encoder.

Unlike a traditional TTS system, these models sometimes do not include a linguistic front-end (Wang et al., 2017). The removal of the front-end component is touted as a benefit because one of the main issues in modular TTS systems is that each of the various modules responsible for different tasks requires feature engineering and often

domain-specific knowledge (Wang et al., 2017; Tachibana et al., 2018). Further, having multiple modules means that errors in one module of the system propagate to subsequent modules (Wang et al., 2017). However, one of the consequences of removing models which specify duration and intonation is that prosodic features are now learned implicitly by the model and therefore cannot be modified at synthesis time, i.e. we cannot control these features independently (Ren et al., 2019). A further consequence of this is that models based on attention which did not explicitly model duration, such as the Tacotron model (Wang et al., 2017), are prone to missing words or *babbling* due to failures in attention (Ren et al., 2019). More recent approaches have therefore introduced methods to condition the prediction of the mel spectrogram on prosodic features such as duration, $F_0$ and intensity, while simultaneously training a model to predict these features along with the mel spectrogram. For example, by conditioning the model on duration, these models can regulate the duration of phones, mitigating the babbling issue. By conditioning on other prosodic features, the model can use information beyond the input specification to predict the mel spectrogram which should lead to better synthesis quality. A by-product of conditioning the model on these features is that these features can then be changed at synthesis time allows the model to be controlled (see Section 2.3.3).

### 2.3.2 Speech Synthesis Models

In this section, we will give an overview of the speech synthesis models which will be used in this thesis. Note that in the course of this thesis, newer models have been developed which show improved performance over the models below. These models were chosen at the time because they were relatively fast to train, could be trained given our computer constraints, and at the time produced reasonable synthesis quality.

#### Deep Convolutional TTS (DCTTS)

DCTTS (Tachibana et al., 2018), is an end-to-end speech synthesis model which employs an attention-based sequence-to-sequence approach using convolutional neural networks (CNNs) in place of the recurrent neural network-based model which was used in Tacotron (Wang et al., 2017). CNNs offer an advantage over recurrent neural networks (RNNs) because the outputs from a single convolutional layer can be computed in parallel while still capturing sequential dependencies, in contrast to the RNN in which each output is computed sequentially. This makes them faster to train than RNN models. DCTTS takes as input a sequence of text which is encoded using a fully CNN-based text encoder, while the mel spectrogram is encoded using fully CNN-based audio encoder. Attention is calculated between the output of the text encoder and the audio encoder to capture the relationship between the text and audio. An Audio Decoder, again based on CNNs estimates the resulting mel spectrogram. At inference time, the speech is generated autoregressively using a causal CNN.

#### FastPitch

FastPitch (Łańcucki, 2021) is a non-autoregressive speech synthesis model based on the transformer architecture. This model is based upon the FastSpeech architecture introduced by Ren et al. (2019). These models marked an improvement on end-to-end systems such as

Tacotron (Wang et al., 2017), which suffered from issues resulting from attention failures, such as babbling.

Like DCTTS, FastPitch does not contain a linguistic front-end as specified in the previous section. However the publicly-available implementation[3] does contain a dictionary look-up based G2P function and some rudimentary text normalisation. The main architecture of FastPitch is comprised of two feed-forward transformer (FFT) stacks which can be seen in Figure 2.2. The first FFT stack functions as the encoder of the model and creates a 384 dimension hidden representation of the input sequence (graphemes or phones). The output hidden representation of the encoder is then fed to three different variance adapters. The variance adapters function to provide prosodic controllability, by conditioning the model on duration per input symbol and the average $F_0$ over the input symbol. By explicitly modelling duration in this way, FastPitch alleviates the durational issues which Tacotron had. In Łańcucki (2021), two variance adapters are described, but in the implementation used in this thesis, an additional variance adapter is used which additionally conditions the model on intensity.

The pitch prediction model takes as input the hidden representation output of the encoder network. The task of this model is to predict the average $F_0$ across each input symbol. In the implementation used in this thesis, $F_0$ is extracted using the librosa (McFee et al., 2023) implementation of the pyin pitch tracking algorithm (Mauch and Dixon, 2014). $F_0$ is standardised using the corpus-wide speaker mean and standard deviation. The pitch prediction model consists of a stack of 1-D convolutional layers which predict the mean $F_0$ per input symbol. The predicted pitch values are then projected to match the dimensionality of the hidden representation (384) using a further 1-D convolution and then summed to the encoder output. The intensity prediction model shares an identical structure to the pitch prediction model, however instead of predicting the mean $F_0$ per input symbol, it predicts the mean intensity per input symbol.

Similarly, the duration model predicts the number of frames per input symbol. Again, this model shares an identical structure to the pitch predictor, but this time the duration values are not projected to match the dimensions of the encoder output as these values are not summed to the encoder output. Here, the duration values are used to upsample the encoder output, which is equal to the length in input symbols, to the number of frames in the mel spectrogram to be predicted. The upsampled encoder output is then passed to the second FFT block which functions as the decoder. The decoder then predicts a sequence of mel spectrogram frames. Each of the models uses a mean squared error loss in predicting the prosodic parameters and the mel spectrogram frames.

During training, the model uses the ground truth conditioning values i.e. it does not utilise the values for pitch, intensity and duration which are predicted by the variance adapters during training, but uses the real values as extracted from the data. At synthesis time, there are two ways in which this model can be used. In the first way, the input text is processed and fed to the encoder, subsequently based on the encoder representation the variance adapters will predict the values for $F_0$, duration and intensity. The other option provides controllability. To control the model, a user can specify values of $F_0$, intensity, or duration and these values will be used instead of the predicted values. After the mel spectrogram is predicted, it is then fed to a neural vocoder which synthesises the final

[3] `https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/SpeechSynthesis/FastPitch`

speech. In this thesis, in Chapter 3 and Chapter 4 we use the WaveRNN (Kalchbrenner et al., 2018) vocoder, and in Chapter 5-7 we use the more recent HiFi-GAN vocoder (Kong et al., 2020).

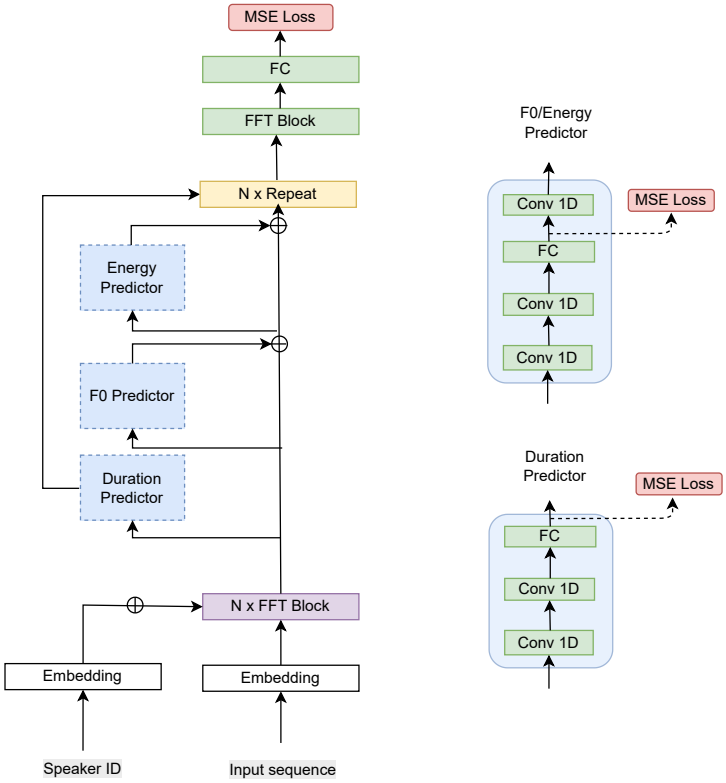


Figure 2.2: *Baseline FastPitch multi-speaker model (adapted from Łańcucki (2021))*

### 2.3.3 Controllability in Speech Synthesis

Controllability within the context of TTS refers to the ability of the model to change certain characteristics of the synthesised speech at inference time. In the context of end-to-end synthesis, this is usually achieved by *conditioning* a model, whereby additional information about the speech to be synthesised is provided to the model, or learned jointly during training. The main idea here is that this additional information should be helpful to the model when predicting acoustic features, such as the mel spectrogram. During training, the model learns how this additional information relates to the mel spectrogram representation. Of course, the success of conditioning is determined by whether there are systematic acoustic correlates in the data that are explained by the conditioning feature.

The ability to condition a model at the input stage is one of the benefits of end-to-end synthesis, in contrast to a traditional TTS system where conditioning each sub-module of the front-end is not straight-forward (Wang et al., 2017). Some spoken characteristics which have been explored within controllability of end-to-end systems include controlling the accent of a speaker, the speaker identity, the prosodic realisation of an utterance, (Suni

et al., 2020; Raitio et al., 2020; Lameris et al., 2022), or the speaking style (Skerry-Ryan et al., 2018).

Controllability can be achieved using different methods and representations, and these methods will be explored in more detail in Chapter 4. For example, in the model FastPitch introduced above, controllability is possible by explicitly conditioning a model on the ground-truth mean $F_0$ and intensity per phone, and duration. Control via other prosodic features, such as global $F_0$, global $F_0$ range and voice quality features (Raitio et al., 2020, 2022b), is also possible. For features such as speaker identity or accent, we can control a model by conditioning the model on a speaker or accent embedding. These embeddings can either be learned from scratch during training, or can be provided using an external embedding, for example a speaker embedding from a speaker verification model (Stan and O'Mahony, 2023).

One of the benefits of adding controllability to a model for conversational speech synthesis is that we can potentially control the model at inference using an external model that has been trained on a larger quantity data than that seen in training, or on data which itself is not suitable for speech synthesis modelling (Raitio et al., 2020; Ben-David and Shechtman, 2021). For example, Ben-David and Shechtman (2021) extracted prosodic features from Switchboard Corpus I (Godfrey and Holliman, 1993) to train an external prosody prediction model which could be used to steer a controllable TTS model. Secondly, the ability to control a speech synthesis model enables the creation of custom listening test stimuli for speech science research. In particular, for studies into conversational speech, creating controlled but conversational-sounding speech is difficult. Thus TTS can allow us to both keep certain features constant, for example the speaker identity and the text, while modifying certain prosodic features, for example turn-taking behaviour (see Chapter 7). In this thesis we provide controllability by using stylised representations of the $F_0$ contour and categorical prominence and boundaries in Chapter 5, and by using a categorical attribute of turn-position in Chapter 7.

### 2.3.4 Conversational Speech Synthesis

#### Acquiring Data for Conversational Speech Synthesis

As we have mentioned, traditional speech datasets are comprised of carefully controlled isolated utterances and are therefore devoid of the communicative characteristics found in conversational speech (Campbell, 2006). To synthesise conversational speech, other sources of data have been used, including parallel datasets that contain both conversational speech and read speech from the same speaker, datasets of reenacted conversation, as well as *found conversational data*.

Andersson et al. (2010), for example, recorded both a set of phonetically-balanced read prompts and spontaneous conversational speech from the same speaker. In their analysis of the prosodic differences between the conversational utterances and the read speech utterances in Andersson et al. (2012), they found no difference in $F_0$ range or global $F_0$ distribution between the two modalities. However, they found that conversational utterances showed more variability in the final $F_0$, than in the carefully recorded read speech utterances, which they attributed to conversation containing a wider variety of speech acts (Andersson et al., 2012).

Other approaches use re-enacted conversational speech (Zandie et al., 2021; Adigwe

and Klabbers, 2022). This approach can be useful because, as in the case of read speech, it provides a level of control over the content of utterances. While this approach offers more control, it results in fewer spontaneous behaviours, such as the reduction in the number of fillers (Adigwe and Klabbers, 2022), overlaps and false starts.

Finally, found data has been used to train conversational models as a source of naturally occurring conversational data (Székely et al., 2019b). As we saw in Chapter 1, found data is data which has not been recorded for the purposes of training speech synthesis models (Saeki et al., 2024), and as mentioned, common sources of this data for conversation include podcasts (Székely et al., 2019b; Chen et al., 2023), audio from YouTube videos (Chen et al., 2023), and existing conversational corpora which were created to train ASR models (Ben-David and Shechtman, 2021). In Chapter 4, we give a more detailed overview of the data used in previous studies of conversational speech synthesis.

### Synthesising Conversational Speech in Context

In Section 2.2.2, we saw that prosody serves many functions in conversation and that the prosodic realisation of an utterances varies as a function of context. Crucially, there are many contextual factors which will exert influence on the realisation of an utterance and these are expressed simultaneously, affecting the same shared prosodic signal in different ways. In this section, we give a brief overview of some of the approaches to incorporating conversational context into TTS models.

Some of the earliest work using context in TTS applied rule-based frameworks to select the appropriate intonational representation given a dialogue context and its information structure (IS) (e.g. Prevost and Steedman (1994); Kruijff-Korbayová et al. (2003)). For example, Kruijff-Korbayová et al. (2003) modelled intonation in context for task-specific dialogue using a rule-based system which analyses the IS of an incoming utterance from a user and generates a response utterance, which is also labelled for IS, and is sent to the TTS system. To adapt the intonation of the response utterance, the proposed method assigns intonational mark-up (GToBI and SABLE) based on the information structural roles of the tokens as specified by the model. During evaluation they compared the default realisation of the TTS model to their proposed realisation, which uses context. They presented listeners with dialogue extracts consisting of three to five turns in which the final turn was provided by their system or the baseline. Crucially, in their evaluation they selected particular contexts which should elicit different information structural realisations of the target utterances. Context utterances were presented in written form, while the target was speech synthesis. Using a rating scale (1-5), participants rated the *contextual appropriateness* of the utterance (Kruijff-Korbayová et al., 2003). They found that the context-aware realisation controlled with GToBI mark-up was rated more highly than the baseline receiving an average rating of 3.71 compared to 3.47 for the baseline system[4].

While the approach of Kruijff-Korbayová et al. (2003) uses a formal system to model context in a constrained dialogue setting, more recent approaches to conversational synthesis have started to explore the addition of the linguistic and acoustic context of the previous utterances in end-to-end TTS systems (e.g. Guo et al. (2021); Yamazaki et al. (2021); Li et al. (2022); Lee et al. (2023); Hu et al. (2022); Cong et al. (2021); Nishimura et al. (2022)). The goal in these approaches is to condition the synthesis of a target utterance on the textual

[4]These results are from the second experiment comparing the baseline to GToBI mark-up. The first experiment compared the baseline (mean rating 3.23), GToBI mark-up (3.62) and SABLE mark-up (3.19)

content or acoustic content of the previous utterance(s) to implicitly learn the contextual relationship between the context and target utterance to improve the prosodic realisation of the target. The number of utterances used in the context window varies considerably across studies, but windows of up to 10 prior utterances have been attested (Guo et al., 2021).

One of the first to use conversational context conditioning in an end-to-end TTS model was Guo et al. (2021). They used the Tacotron 2 (Shen et al., 2018) model to synthesise contextualised Chinese conversational speech trained with semi-scripted dialogue from voice-agent interactions. To model context, they added two context encoders. First they used an auxiliary encoder to capture information about the textual content of the target turn, which comprised of one or more utterances. The input to the auxiliary encoder was token-level BERT embeddings and various character- and utterance-level positional information about each utterance in a turn. Second, to capture the previous linguistic context, they included a conversational context encoder, this time using a sentence-level BERT embedding to encode each previous turn. They informally evaluated that a context window of 10 turns led to better performance. To evaluate the model, listeners heard five entire conversations between a male speaker and the proposed TTS system. Listeners took part in a comparative mean opinion score experiment (CMOS) and rated each turn from the proposed TTS system and baseline, after which they listened to the entire conversation and rated overall impression. They found that encoding the previous context was preferred to only encoding the target utterance information receiving a CMOS score of 0.18 when rating each utterance and 0.39 for the overall rating of the conversation.

Yamazaki et al. (2021) used the previous textual and acoustic context to train both an $F_0$ prediction network and a dialogue response generation system[5] for Japanese conversation, which predicted a system response following a user utterance. Specifically, their goal was to model entrainment, by using the previous acoustic realisation of the user to facilitate the prediction of the $F_0$ from the system. The $F_0$ prediction model is based on an LSTM sequence-to-sequence model which takes as input the response to be generated as well as the previous utterance and its $F_0$ realisation per token. The output of the system was the $F_0$ information per token of the target utterance, which was then used to control a TTS model which was conditioned on the $F_0$ information. They used a CMOS evaluation in which two systems were compared and in which each target utterance was heard in context. They found that the model which included the previous prosodic and lexical context was significantly more preferred over the baseline system and the model which did not use the previous context but which generated $F_0$ using the current utterance only.

Finally, to model both the previous linguistic context of the conversation and the previous prosodic context, Li et al. (2022) used a multi-modal representation of the prior five utterances in the conversational context, encoded in a graph-based network in which every node in the graph represents the prosodic and textual representations of the previous utterances. Each utterance is represented by a BERT embedding which is passed through an additional encoder and the acoustic realisation of the previous utterances is passed reference encoder and Global Style Token (GST) module (Skerry-Ryan et al., 2018). The graph-based neural network, which models the sequential dependencies between the context utterances, creates a richer representation of the context. An attention layer is used

[5] They also trained a network which predicted both the tokens to be realised and the associated $F_0$ of those tokens, but the separated model performed better.

to capture the importance of the contextual features in relation to the current utterance. The output of this module is a set of predicted GST weights which specify the style of the target utterance. These are then passed to a FastSpeech 2 Ren et al. (2021) model, which is additionally conditioned with a GST module to control the prosodic realisation of the target utterance. To train the model, they used YouTube videos demonstrating conversation for L2 Learners of English. For evaluation they used conversational segments comprising of six utterances, of which the final utterance is the target and performed a MOS and ABX preference test. The comparison took place between a baseline (a model conditioned on the linguistic content of the previous utterances, following Guo et al. (2021)) and their multi-modal approach. They found that modelling both the previous prosodic content and the previous text led to higher ratings of utterances (3.356 for the baseline 3.584 for the context-aware system) and a higher percentage of preferences (55.11% preference for the context-aware system compared to 26.22% for the baseline[6]).

As we have seen, there are a number of approaches which have been used to model context (Hu et al., 2022; Cong et al., 2021; Nishimura et al., 2022), however there are many limitations of these approaches. For example, the work of Kruijff-Korbayová et al. (2003) relies on a rule-based system and therefore can only be applied to very limited and possibly contrived dialogue settings. It is therefore unlikely that this can be used as a system to model open-domain casual conversation. The other approaches have the advantage that they can be applied to any conversation, however one of the main issues with these approaches lies in their evaluation. First, the specification of a context window, which in these studies ranges from one utterance in the past to ten utterances in the past seems arbitrary and is not based on empirical evidence. Thus, ablation studies are needed which vary these context windows to empirically investigate how much context is needed. Second, as context length increases, the likelihood of that specific context occurring, or generalising across other examples, in a relatively small corpus is small. This means that the context could be functioning as a unique embedding for that utterance which captures all of the variance in the acoustic signal, but is not itself capturing context (similar to Watts et al. (2015)). To test this, it would be beneficial to train a baseline model with random contexts and evaluate the model with random context utterances to ensure that the model is indeed learning real context effects.

Lastly, we have seen that there are many contextual factors which can affect an utterance, but the evaluation used in the work on end-to-end synthesis above does not seek to evaluate any particular contextual effect, thus we do not know what the contextual representations do and don't capture. The work from Kruijff-Korbayová et al. (2003), on the other hand, though only focusing on a limited contextual effect, chooses to evaluate dialogue segments in which they know certain information structural effects should be elicited. In this thesis, we therefore take a controlled approach to context, by focusing on very targeted contextual effects. In Chapter 6, we investigate the role of prosody in the perception of the stance of agreement/disagreement, while in Chapter 7 we focus on the prosodic cues in turn taking and create stimuli which are textually turn-ambiguous.

[6]This does not include 14.67% no preference and it is unclear why the preference scores do not sum to 100%.

## 2.4 Speech Synthesis Evaluation

As we have seen, the traditional paradigm when training a TTS model has been to synthesise utterances in isolation. Because of this, TTS evaluation has also been developed for isolated utterances. In the following section, we will give an overview of evaluation methods which will be featured in this thesis. This section only serves to give a brief overview of these paradigms, a more comprehensive review will be given in Chapter 3, where we will also discuss evaluation in context.

### 2.4.1 Mean Opinion Score (MOS)

A MOS test involves participants being presented with speech synthesis output from two or more systems. The task of the participant is to rate each utterance (usually one by one), using a Likert scale which describes a particular property, such as the degree of naturalness or intelligibility of the speech. Usually this scale is between one and five, however higher scales and incremental scales have also been attested (e.g. Clark et al. (2019); Kirkland et al. (2023)).

The MOS paradigm has recently been subject of critique (e.g. Wagner et al. (2019) and see Cooper et al. (2024) for discussion). For example, it may not be the most suitable paradigm for evaluation of contextualised speech or for evaluating speech as the quality continues to reach high levels. We return to this point in Chapter 3. Further, when rating utterances using ill-defined terms, such as *naturalness*, it may not be clear which exact dimensions of the speech signal listeners are attending to and MOS cannot easily diagnose which aspects of the speech listeners are using to make their final judgement (Gutierrez et al., 2021). Finally, analysing MOS scores can be difficult due to the ordinal nature of the rating and the fact that participants may use the Likert scale differently. For example, a listener may rate system A as one and system B as two, while another listener rates system A as three and system B as four. In this thesis, we address the scale calibration issue by using ordinal mixed effects models (Christensen, 2022).

Though the MOS paradigm has faced critique, we must also view this paradigm in the context in which it arose. Speech synthesis quality has not always achieved the levels of naturalness that it can now achieve. Even with the recent advances in TTS modelling, using data with high amounts of variation, such as conversational found data, can lead to unnatural and sometimes unintelligible speech synthesis output. Thus the MOS test can therefore be used a valid pre-test to examine the general quality of the speech output.

### 2.4.2 Preference Tests

In this thesis, next to MOS tests, we also use pairwise preference tests to compare systems. In this paradigm, participants are presented with the speech synthesis output of two systems and the task this time is to make a side-by-side comparison. The nature of the comparison varies across studies. In the simplest form, participants may be asked which rendition they *prefer*. However, in Chapter 4, we ask participants which rendition sounds more *conversational* and in Chapter 5 we present listeners with a reference recording and ask which of the two renditions sounds most like the reference. The form that the comparison takes can be either a forced-choice preference task in which the participant simply picks system A or system B as their preferred system (or in some cases a third option

of *no preference* is allowed), or it can take the form of a comparative mean opinion score (CMOS) test in which the participants are asked to specify the degree to which they prefer one system over another using a scale (Loizou, 2011). In this thesis we exclusively used forced-choice binary tasks.

One advantage of using preference tests is that they allow for participants to directly compare the quality two speech synthesis systems. For example, Camp et al. (2023) found that CMOS preference tests were more suitable for system comparison and more reliable when fewer participants could be recruited compared to MOS tests. It is unclear, however, the degree to which participants in a CMOS paradigm utilise the rating scale differently – in the study of Camp et al. (2023) the CMOS scale ranged from -3 to 3. We therefore may face the same scale calibration issues as those mentioned in the previous section. This motivates our use of a forced-choice binary task which can mitigate the scale calibration issues that MOS results can have, providing a more simple task for participants.

# Part II

# Evaluation in Context

# 3

# Evaluation of Synthetic Speech in Context

This chapter is based on the following paper:

**O'Mahony, J.**, Oplustil Gallegos, P., Lai, C., & King, S. (2021). Factors Affecting the Evaluation of Synthetic Speech in Context. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)* 148-153. doi:10.21437/SSW.2021-26

Author Contributions: The first author was responsible for developing the methods for each of the experiments, creating the textual stimuli, training the speech synthesis models, creating the prosodic variants of the stimuli, carrying out the listening tests, analysing the results of the listening tests and writing the main draft of the paper above. The contribution of the second author was in also developing the research questions and method. The third and forth authors' roles were supervisory in nature and both supervisors contributed to editing and feedback on the published paper which has significant textual overlap with this chapter.

Additionally, the synthetic speech was post-processed to remove noise by Carol Chermaz and the training data text data with prosodic labels was provided by Antii Suni.

## 3.1 Introduction

In Chapter 2, we saw that there has been a recent shift in TTS research from synthesising single utterances to synthesising long-form content, for example, a multi-utterance turn in a conversation, or a paragraph from a book. The simplest approach to synthesising long-form content, and one which is still predominantly used, is to synthesise multiple utterances in isolation and subsequently concatenate the isolated utterances together to

form a longer unit. The crucial issue with this approach, however, is that we do not speak in isolated utterances. In conversation, the topic of this thesis, but also in monologue data, such as audiobook narration, each utterance is said in a particular *context*.

Recent improvements in TTS modelling have paved the way for approaches which can take certain aspects of context into account. As we saw in the previous chapter, there are many contextual factors which exert influence on how something is said, though many factors will differ between long-form monologue data and conversational data. Accounting for context has the potential to capture long-range text dependencies, such as discourse (Aubin et al., 2019) and paragraph information (Farrús et al., 2016), which are known to affect the prosodic realisation of an utterance. Context-sensitive prosody should exhibit increased prosodic variation compared to *default* prosody generated for isolated sentences, for example by accounting for information structural dependencies (Prevost and Steedman, 1994), and thus better long-form TTS (Hirschberg, 1990; Prevost and Steedman, 1994). At the time of writing this thesis, most work has modelled contextual features which can directly be derived from long-form data, for example using previous textual content (Guo et al., 2021; Xu et al., 2021) and/or acoustic content (Oplustil-Gallegos and King, 2020; Oplustil-Gallegos et al., 2021) of the preceding utterance(s). These context-aware models should generate different prosodic renditions of the same textual content depending on which previous context utterance(s) they are conditioned on at inference. However, a key question then arises: how do we evaluate the resulting renditions?

As mentioned, TTS output has been almost exclusively generated utterance-by-utterance and has therefore also been evaluated using isolated utterances (Clark et al., 2019; Wagner et al., 2019). For context-sensitive TTS, appropriate evaluation paradigms are not yet fully developed. One difficulty when rating prosodic variability is that countless realisations may be equally valid given a specific context (Wagner et al., 2019). Further, rating an utterance in context is a fundamentally different task to rating an utterance in isolation: varying the context can change the rating of the utterance. Conversely, in isolation the listener does not have access to any contextual information (Clark et al., 2019) (although participants might be able to imagine it (Hodari et al., 2020; Adigwe et al., 2024)). This could potentially cause marked prosodic forms, elicited by a very specific context, to be rated lower when presented out of context, where listeners would expect default prosody. The opposite could also be true: perfectly natural and well-spoken utterances are rated highly in isolation, but when heard in an infelicitous context they are rated lower. Unfortunately work on speech perception in context, in both speech science and speech synthesis evaluation, is sparse.

Clark et al. (2019) were the first to study the impact of context on the ratings of long-form speech synthesis. They found that utterances presented in isolation received a significantly different MOS than utterances presented in context, with those heard in context receiving a higher rating when both the context and target were synthetic speech. Importantly, the synthetic speech used in their study was *not* context-sensitive. This boost in MOS calls into question whether the MOS paradigm is the correct approach for evaluating synthetic speech in context. In this thesis, this question is central because the found data that will be used throughout this thesis was spoken in context, and because we are primarily concerned with using this data to ask why speakers choose certain prosodic forms. It is therefore imperative that a suitable evaluation paradigm is chosen.

The goal of this chapter is therefore to assess whether MOS is a suitable experimental

paradigm for evaluating speech in context and for use as an experimental framework in this thesis. More specifically, the goal is to investigate what factors led to the differences in MOS ratings found by Clark et al. (2019). We conducted three experiments to answer the following research questions:

**RQ3.1** Do the **instructions** have an effect on MOS ratings of utterances presented in context?

**RQ3.2** Does between-sentence **textual context dependency** have an effect on MOS ratings?

**RQ3.3** Is the MOS paradigm suitable for rating **prosodically-varied** synthetic speech?

Although Clark et al. tested a range of presentation types, including paragraphs, we will focus on a comparison between isolated utterances and context-target pairs in which a target utterance is presented after a single context utterance. Finally, though this thesis is predominantly concerned with conversational speech, in conversation the prosodic realisation changes as a function of much more than the preceding sentence, e.g., pragmatic context, emotional state of the speaker, speaking partner, genre of speech etc. (Wagner et al., 2019). To avoid extraneous confounding factors, such as the speaking partner of the preceding utterance, as well as other pragmatic factors, we will concentrate on prosodic realisations which are determined by the textual context alone. For this reason, we are using read speech as opposed to conversational speech [1]. In the next section, we will describe research that had been published *before* this study was conducted. However, in the discussion we will present some more recent work on evaluation by others and discuss their findings in light of the results found in this chapter.

## 3.2 Related Work

### 3.2.1 Evaluation in context

As TTS synthesis approaches its limit of naturalness for isolated utterances, there is an increasing focus on prosodic variability (Hodari et al., 2020; Klimkov et al., 2017; Tyagi et al., 2020, for example) including the use of surrounding context to condition the realisation of the current utterance (Guo et al., 2021; Xu et al., 2021; Oplustil-Gallegos and King, 2020; Oplustil-Gallegos et al., 2021). The main focus of TTS evaluation historically, as we saw in Chapter 2, has been to compare the naturalness and intelligibility of the speech output of two or more systems. But as we move towards a goal of prosodic appropriateness, the focus of evaluation needs to shift. We can imagine that a system, which can produce prosodically-varied renditions of an utterance, might produce speech which sounds equally natural and intelligible, but that each rendition might be more suitable in a particular linguistic or pragmatic context. There is, however, little agreement on the best method for evaluating such prosodically-varied synthetic speech.

Some opt to use a qualitative approach. After testing whether prosodic realisations were perceptually distinct using a discriminative task, Hodari et al. asked participants to

[1] Note that the decision to use read speech was also a consequence of the lack of availability of conversational data which is suitable to train speech synthesis models. This will be addressed in Chapter 4 and 5.

judge what effect different prosodic renditions had on the interpretation of the sentence, e.g., subtle differences in meaning or intent (Hodari et al., 2020). They found that participants were able to describe different contexts or situations where the prosodic variant would be found. A different qualitative approach was taken by Xu et al. (2021) who constructed different textual contexts and used these to generate different prosodic realisations of a single sentence in order to determine what effect their BERT-based context-aware model had on the prosody of a sentence (Xu et al., 2021).

Others opt for quantitative subjective evaluation using a MUSHRA-like paradigm. For example, Tyagi et al. used linguistic information, such as syntactic information and word embeddings to generate richer prosodic variability and evaluated both isolated utterances and long-form material (Tyagi et al., 2020). In order to assess the quality of the prosodic output of individual sentences, they asked ten linguists to judge the appropriateness of the prosody in isolation. They claimed that judging prosody requires domain-specific knowledge. This raises an issue with devising appropriate metrics for prosodic felicity, if using non-expert listeners requires them to have an awareness of this dimension of the speech signal. Even among experts, however, it has been shown that inter-annotator agreement can be quite low (Syrdal and McGory, 2000). For long-form evaluation, Tyagi et al. (2020) used crowd-sourced listeners and asked them to rate whole news stories for the "*suitability*" (Tyagi et al., 2020, p.4409) of the speaker's style, which they said would assess naturalness. As we will see from the results of Experiment 1 (Section 3.4.3), changing just one word in the task instructions can lead to different ratings. Many studies have used instructions such as *suitability* and *appropriateness* as synonyms for naturalness when they are in fact asking something quite different (Clark et al., 2019; Tyagi et al., 2020).

Another option is a preference test to determine which system or prosodic realisation listeners prefer. For example, Aubin et al. (2019) tested the difference between a TTS system using discourse relations and a baseline system, using a preference test in which the target utterance was presented in a natural speech context. Oplustil-Gallegos and King (2020) also used a preference test in order to evaluate whether systems that take acoustic context into account from the preceding sentence perform better than a non-context-aware baseline. While preference tests and MUSHRA both ask participants to make *direct comparisons* of stimuli with differing prosodic renditions, MOS tests do not. By asking listeners to provide *absolute* ratings, many stimuli could receive the same MOS score.

To evaluate the use of MOS for long-form evaluation, Clark et al. (2019) compared differences in MOS ratings for utterances presented in isolation, in a context-target pair, or in a paragraph. They asked participants to rate the *naturalness* of utterances presented in isolation, but for context-target pairs, they asked participants to rate *appropriateness* of the target utterance given the context. The type of context was also varied, being either text, synthetic speech, or natural speech. They found that target utterances presented in context were rated significantly higher than the same utterances presented in isolation, when the context was in the form of text or synthetic speech. It is important to re-iterate that the synthetic speech was *not* context-dependent.

Clark et al. (2019) postulated that the increase in MOS rating in context might be due to the task specification, recall that the wording of the instructions differed between the evaluation of isolated utterances and in-context utterances. Previous work has indeed found that instructions can have an impact on MOS rating (Dall et al., 2014b). Clark et al. (2019) also suggested that the increase in MOS ratings may be due to "the fact that the content of a paragraph non-initial sentence sounds less natural when presented out of

context." (Clark et al., 2019, Section 5.1). They found no increase in MOS ratings when the preceding context utterance was (non-vocoded) natural speech, reasoning that mismatches in quality between natural and synthetic speech make the synthetic speech sound of lower quality.

In the study reported in this chapter, we focus exclusively on the MOS paradigm and investigate what factors lead to differing MOS scores between utterances presented in isolation and utterances presented in context. Clark et al. (2019) inadvertently used different wording of instructions when presenting isolated utterances than when presenting them in context. One of our experiments investigates the effect of wording alone, to avoid this confound. We restrict the investigation to the case of both target and context being synthetic speech. We also investigate whether the paradigm is sufficiently sensitive to differentiate prosodically-different renditions of a sentence by a single system, something that Clark et al. (2019) did not do. This question is vital because it would be useful to evaluate systems that can output multiple renditions of an utterance, for example when testing the difference between conditioning the realisation of an utterance on its true context or a random context (Oplustil-Gallegos and King, 2020).

## 3.3 Research Questions

### 3.3.1 Effect of instructions

As noted in Clark et al. (2019), the increase in MOS rating between the isolated condition and the TTS context condition was rather unexpected, given that the TTS model in question was not context-aware. One factor that might have influenced MOS ratings was the inadvertent different wording used between the in-context condition and the isolated condition. Specifically, participants were asked to rate the *naturalness* of isolated utterances but the *appropriateness* of utterances presented in context. By wording the instructions to ask for either naturalness or appropriateness ratings, Experiment 1 investigates whether this difference leads to changes in rating, independent of how the stimuli are presented.

### 3.3.2 Effect of between-sentence textual context-dependency

Although Clark et al. (2019) suggested that the increase in MOS rating may have been due to the task, they also suggested that utterances from non-paragraph-initial position may benefit from being presented with a preceding context. This is because non-initial sentences more often contain anaphoric references, such as pronouns, and therefore need a context in order to be fully understood. In Experiment 2, we manipulate the context-dependency of the target sentence text to test whether sentences containing anaphora receive higher MOS ratings when presented in a context that provides the referent, than non-anaphoric versions that do not need context in order to be fully understood.

### 3.3.3 Sensitivity of MOS to prosodic differences

While Clark et al. (2019) investigated the effect of synthetic spoken context, natural spoken context, and text context, they did not investigate whether participants are sensitive to changes in prosodic realisation when both context and target are synthetic and differ only

in their prosody. Both Wagner et al. (2019) and Tyagi et al. (2020) suggest that rating speech in context is difficult because there is no *correct* realisation and multiple variations will be equally acceptable. Therefore, in Experiment 3, we make one stimulus obviously non-canonical and ill-fitting to the context, in order to evaluate whether such a mismatch is salient for participants. If participants rate both the non-canonical and canonical highly in context, that would be evidence that this task is ill-suited to evaluating prosodic variation.

## 3.4 Methods

### 3.4.1 Data and models

We used the LJ Speech dataset, a corpus of audiobook data, which consists of roughly 13 000 sentences read by a female speaker (Ito and Johnson, 2017), for training all models. The model used in all experiments was the Ophelia implementation[2] of DC-TTS (Tachibana et al., 2018), which was described in Chapter 2. All stimuli were vocoded with a pretrained WaveRNN (Kalchbrenner et al., 2018) trained on LJSpeech for 800k steps[3].

For Experiment 3, we need to manipulate prosody. We used the publicly-available training data used in Suni et al. (2020), which comprises the LJ Speech dataset marked up with prominence and boundary labels automatically generated using the Wavelet Prosody Toolkit[4]. This toolkit uses a CWT over the combined signals of $F_0$, intensity, and duration and returns a prominence and boundary strength for each word. These are then discretised to create labels and used as prosodic mark-up in the input sequence. When training our TTS model, the input to the model is now not just a phone sequence, but a phone sequence with the additional prosodic mark-up. Here each word in the utterance receives mark-up which describes whether a work receives a pitch accent or prosodic boundary. This mark-up is learned by the model, just as the phone information is learned by the model. This method therefore provides a method of controlling the prosodic realisation at inference time. By simply by changing the prosodic labels of the input text, a different prosodic rendition can be generated.

Suni et al. (2020) used three strength levels for both the accent and the boundary labels, with accent level 0 signifying a deaccented word and boundary level 0 signifying no prosodic boundary, and level 1 signifying accented words and boundary level 1 signifying an intermediate phrase boundary. Level 2 accent signifies an emphasised wordm and level 2 boundary is roughly equivalent to an intonational phrase boundary (Suni et al., 2020).

The LJ Speech recordings contain some background noise and reverberation, which we mitigated by post-processing all generated synthetic speech with the Automatic Sound Engineer (ASE) (Chermaz and King, 2020).

### 3.4.2 Stimuli

We created 110 pairs[5] of sentences each comprising a context sentence followed by a target sentence, using facts from Wikipedia. An example of two context-target pairs is given in

[2] `https://github.com/CSTR-Edinburgh/ophelia`

[3] See `https://github.com/cassiavb/Tacotron`

[4] `https://github.com/asuni/wavelet_prosody_toolkit/`

[5] Stimuli can be found: `https://johannahom.github.io/SSW-samples/index.html`

Table 3.1. The same sentences were used in all experiments. We did not create test material using held out utterances from LJ Speech dataset because this was too restrictive for carefully crafting suitable sentence pairs. All sentences were phonetised using a G2P model (Park, 2019) and were manually corrected, then synthesised.

Stimuli comprising a context-target pair were created by synthesising the two sentences separately then concatenating them into a single audio file separated by a 400 ms pause, a duration chosen through informal listening. This differs from Clark et al. (2019), who asked listeners to click separate buttons to play context and target utterances.

### Text manipulation

Each stimulus is the synthesised speech of a context sentence followed by one of two possible sentences: either the context-dependent follow-up (CD) or context-independent follow-up (CI). Table 3.1 provides an example. The CD target sentence needs the context sentence for the listener to resolve the anaphoric reference, such as *it* or *they*. In the CI condition, the target sentence has the referent filled in. The only difference between CI and CD conditions is the referent. Any two-word referents were matched with a two-word anaphoric reference so that the number of words in both conditions is the same.

| | Condition | |
|---|---|---|
| | Context-dependent | Context-independent |
| Context | **Storms** have been named in the US since the 1700s, for the UK it's a relatively new thing. | **Storms** have been named in the US since the 1700s, for the UK it's a relatively new thing. |
| Target | The first **one** to receive a name in the UK was storm Abigail in 2015. | The first **storm** to receive a name in the UK was storm Abigail in 2015. |

Table 3.1: *Example of context-dependent (left column) and context-independent (right column) sentence pairs.*

### Prosodic manipulation

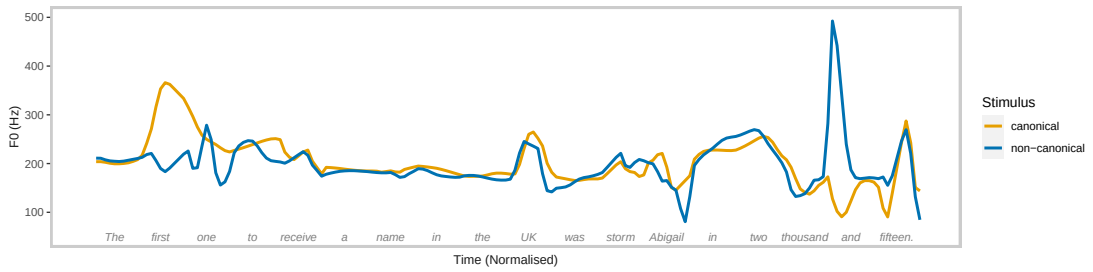


Figure 3.1: *Time-normalised F0 contour of a canonical and non-canonical stimulus from Experiment 3.*

To achieve prosodic manipulation, we manually modified the CWT labels on the input to the TTS model in order to create a *canonical* and a *non-canonical* rendition of each target sentence. Non-canonical renditions (as judged by the first author) were created by changing the accent and phrase boundary structure of the target utterances such that accents were placed on unexpected words (e.g., function words) or placing prosodic phrase boundaries in unexpected places. Figure 3.1 provides an example: *first* is de-accented in the non-canonical rendition, but accented in the canonical rendition; *and* receives a strong emphasis in the non-canonical rendition, but is de-accented in the canonical rendition. The creation of prosodic variants was constrained by the ability of the model, which did not render intelligible speech for every possible combination of accents and boundaries.

Please, read the instructions carefully:
- You will be presented with **one sentence at a time**.
- We want you to rate how **natural** the sentence **sounds**.

(a) *Rating naturalness of utterances presented in isolation*

Please, read the instructions carefully:
- You will listen to **two sentences**.
- The second sentence will be highlighted in **bold text**.
- We want you to rate how **natural** the second sentence *sounds*, given the first sentence.

(b) *Rating naturalness of target utterances presented in context*

Please, read the instructions carefully:
- You will listen to **two sentences**.
- The second sentence will be highlighted in **bold text**.
- We want you to rate how **appropriate** the second sentence *sounds*, given the first sentence.

(c) *Rating appropriateness of target utterances presented in context*

Table 3.2: *Participant instructions.*

### Participants

Listeners who self-reported to have no hearing impairment, be resident in the United States and have English as their first language were recruited through Prolific.[6] No other

[6]https://www.prolific.co

demographic information was asked for. None were allowed to participate more than once within this study. They received monetary compensation for taking part. Participants were asked whether they were using headphones. The responses from anyone who answered *no* were removed from analysis, following Clark et al. (2019), as were those from participants who took less than 10 minutes (the minimum time required to listen to all stimuli).

**MOS task**

We implemented the MOS task in Qualtrics.[7] Following Clark et al. (2019), participants were asked to rate stimuli on a scale of 1-5 in 0.5 increments (a 9-point scale). Points 1 to 5 were labelled as *poor*, *bad*, *fair*, *good*, and *excellent*.

### 3.4.3 Experiment 1 - Effect of instructions

Each participant was assigned to one of 3 conditions. All participants in any given condition rated the same stimuli.
Condition 1: each participant was given the instructions in Table 3.2a, then rated 110 isolated sentences comprising all 55 unique context sentences, and 55 target sentences (a mixture of CI and CD) presented in randomised order. Condition 2: each participant was given the instructions in Table 3.2b then rated 55 context-target pairs presented in a random order. Condition 3: identical to condition 2, except using the instructions in Table 3.2c.

### 3.4.4 Experiment 2 - Effect of between-sentence context-dependency

Each participant rated one of 4 sets of stimuli: Set 1: each participant was given the instructions in Table 3.2a, then rated 110 isolated sentences comprising all 55 unique context sentences, and 55 target sentences (a mixture of CI and CD) presented in randomised order. (*Since this is identical to Experiment 1 condition 1, the same participant responses were re-used.*) Set 2: identical to set 1, and also using all 55 unique context sentences, except now using the remaining 55 target sentences not presented in condition 1 (also a mixture of CI and CD), to counterbalance. Set 3: each participant was given the instructions in Table 3.2c then rated all 55 context-target pairs presented in a random order. (*Since this is identical to Experiment 1 condition 3, the same participant responses were re-used.*) Set 4: identical to set 3, except using the remaining 55 sentence pairs not presented in set 3, to counterbalance.

### 3.4.5 Experiment 3 - Sensitivity of MOS to prosodic differences

Each participant rated one of 4 sets of stimuli: Set 1: each participant was given the instructions in Table 3.2a, then rated 110 isolated sentences comprising all 55 unique context sentences rendered canonically, and 55 target sentences of which around half were rendered canonically and the rest rendered non-canonically, all presented in randomised order. Set 2: identical to set 1, with the same canonical renditions of all 55 unique context sentences, except with the canonical vs. non-canonical renditions of the target sentences swapped, to counterbalance. Set 3: each participant was given the instructions in Table

[7] https://www.qualtrics.com/

3.2c then rated 55 context-target pairs presented in a random order. Context sentences were always rendered canonically. Around half the target sentences were rendered canonically and the rest rendered non-canonically. Set 4: identical to set 3, except with the canonical vs. non-canonical renditions of the target sentences swapped, to counterbalance.

## 3.5 Results

All analyses were done in a by-items fashion such that, for each stimulus, the MOS rating is the mean of all participants' ratings for that stimulus. All data were found to be normally distributed following an insignificant Sharpiro-Wilk test and we therefore used two-tailed paired t-tests. Whenever making multiple pairwise comparisons, p-values were adjusted with Bonferroni coefficients.

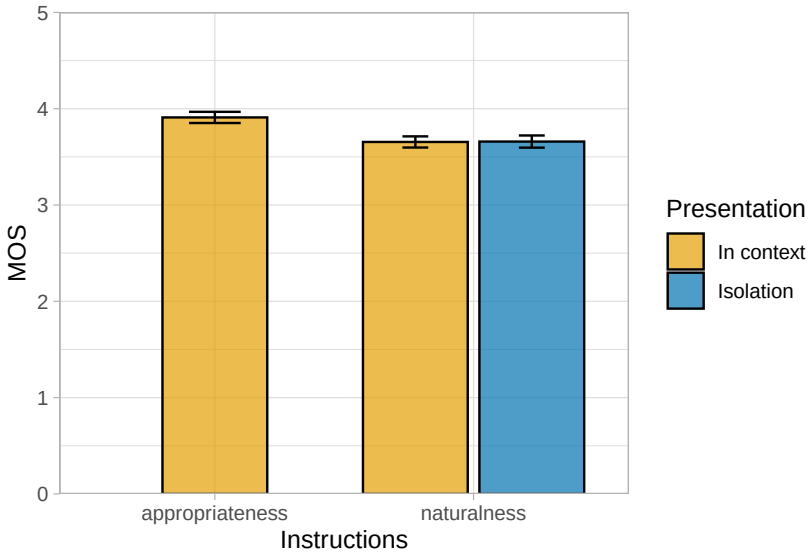


Figure 3.2: *Results for Experiment 1: MOS ratings of appropriateness and naturalness for utterances presented in isolation and in context.*

### 3.5.1 Experiment 1

The experiment tests whether the instruction to listeners affects their ratings. A total of 108 participants took part, of which 8 (7.4%) were removed using exclusion criteria from Section 3.4.2. As we see in Figure 3.2 (blue bar), stimuli were rated lower on the 5-point MOS scale when presented in isolation (M = 3.66 SD = 0.239) than in context. However this is only the case when using the instructions in Table 3.2c which asked them to rate how *appropriate* they sounded in context (M = 3.91 SD = 0.220) but *not* when using the instructions in Table 3.2b which asked them to rate how *natural* they sounded in context (M = 3.65 SD = 0.219). Ratings obtained with the 'how appropriate' instructions were significantly higher than those obtained with the 'how natural' instructions: t(54) = 9.94, $p < 0.001$. When using the 'how natural' instructions, there is no significant difference in ratings for stimuli presented in isolation vs. in context: t(54) = -0.16, $p = 1$. This refutes Clark et al.'s (Clark

et al., 2019) hypothesis that it is the quality of the context and the match in quality (i.e., both context and target are synthetic speech) which leads to an increase in MOS rating.

A better explanation, also mentioned by Clark et al. (2019), is that differences in ratings arise because participants interpret 'appropriate' differently to 'natural'. This implies that, in the condition from Clark et al. (2019) where a synthetic utterance is presented after a natural spoken context utterance, listeners were rating the target as less *appropriate* rather than less natural: it is not appropriate for speech to change from natural to synthetic. We conclude that asking for ratings of *appropriateness* is different to asking for ratings of *naturalness*, for stimuli presented in context.

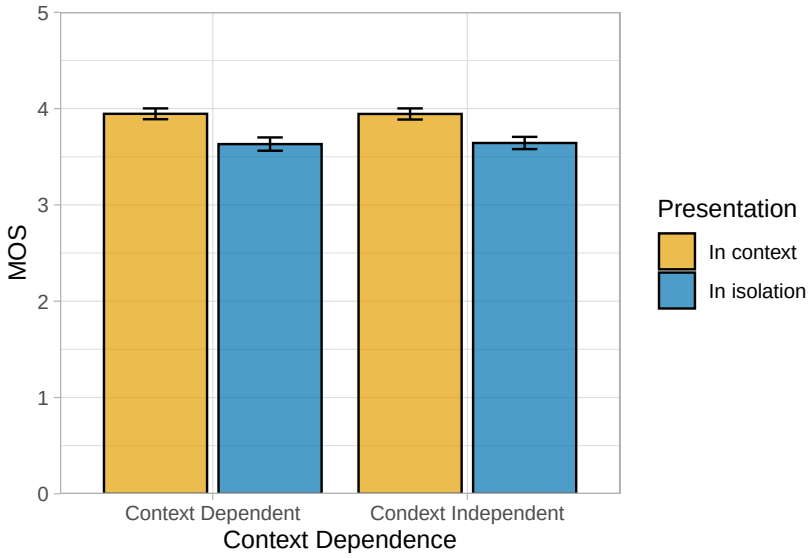


Figure 3.3: *Results for Experiment 2: MOS ratings of context-dependent and context-independent utterances presented in isolation (naturalness) and in context (appropriateness).*

### 3.5.2 Experiment 2

This experiment tests whether ratings of appropriateness are affected by textual dependence between the target and its context. A total of 144 participants took part of which 10 (6.9%) were removed using the exclusion criteria in Section 3.4.2. Results are shown in Figure 3.3. First, for utterances presented in isolation, there is no significant difference in ratings of naturalness for context-dependent (M = 3.63 SD = 0.262) and context-independent (M = 3.64 SD = 0.240) sentences (t(54) = -0.34, $p = 1$). When rated in context, there is no significant difference in ratings of appropriateness between context-dependent (M = 3.95 SD = 0.212) and context-independent utterances (M = 3.94 SD = 0.219): t(54) = 0.048, $p = 1$. Finally, consistent with the results from Experiment 1, there is a significant difference between ratings of isolated utterances and utterances presented in context. This is true regardless of whether the utterance is context-dependent or is context-independent: t(54) = -8.30, $p < 0.001$ and t(54) = -10.48, $p < 0.001$ respectively. We conclude that textual context dependence does not affect listeners' ratings. However, as in Experiment one, ratings of appropriateness for utterances presented in context are higher than ratings of naturalness for utterances presented in isolation.

### 3.5.3 Experiment 3

This experiment tests whether MOS rating is sensitive to differences in prosodic realisation. A total of 144 participants took part of which 13 (9.0%) were removed using the exclusion criteria in Section 3.4.2. Results are shown in Figure 3.4. When presented in isolation, naturalness ratings of non-canonical renditions (M = 3.33, SD = 0.318) were significantly lower than of canonical renditions (M = 3.77 , SD = 0.231), t(54) = 9.41, $p < 0.0001$. This also holds true when these stimuli were presented in context and rated for appropriateness, although ratings of non-canonical (M= 3.86 SD= 0.273) and canonical (M = 4.02 SD = 0.237) are closer: t(54) = 3.86, $p = 0.001$. Both canonical renditions and non-canonical renditions received higher appropriateness ratings when presented in context than naturalness ratings when presented in isolation: t(54) = -7.29, $p < 0.001$ and t(54) = -18.33, $p < 0.001$ respectively. This is consistent with the findings reported in Clark et al. (2019) and our results in experiments 1 and 2. We conclude that MOS is sensitive enough to measure prosodic differences. As in experiments 1 and 2, we once again conclude that appropriateness ratings for utterances presented in context are higher than naturalness ratings for utterances presented in isolation.

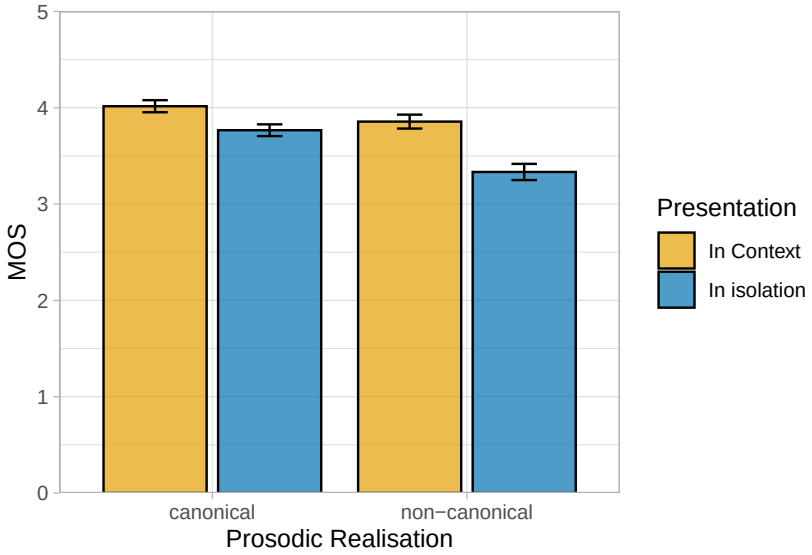


Figure 3.4: *Results for experiment three: MOS ratings for prosodically canonical and non-canonical renditions, presented in isolation and in context.*

## 3.6 Discussion

Like Clark et al. (2019), we found that utterances presented in context receive higher ratings of appropriateness than when presented in isolation, across all three experiments. As we saw from the results of experiment 1, rating how *natural* a given sentence sounded in context did not lead to a significant increase in MOS rating. In experiment 1, we concluded that asking whether an utterance sounds *appropriate* in context is not the same as asking whether it sounds *natural*. We believe the boost in rating is caused by the task specification, as Clark et al. suggested. This could be because the term *appropriate* is open to interpretation by listeners as textual appropriateness or prosodic appropriateness. Given that the stimuli in

both conditions were identical, however, means that the rating of *appropriateness* cannot be fully based on prosodic differences. Additionally, this highlights the importance of reporting the exact wording of instructions given to participants during evaluation (see also Kirkland et al. (2023)).

Clark et al. (2019) hypothesised that target utterances might be rated lower in isolation because the texts contain referential material that is only disambiguated when presented in a context which provides the correct referents. We therefore tested whether context-dependent targets received a boost in rating when their context was provided compared to near-identical context-independent targets in which the referent is provided in its full form. The results from experiment 2 suggest this is not the case. The context-dependency of text did not play a significant role in listeners' ratings as evidenced by the identical increase in MOS ratings for both context-dependent and context-independent target utterances. This does not mean that participants were not taking the text into account at all. All our sentence pairs (an example is in Table 3.1) fitted together contextually, whether the target contained anaphoric reference or not: so all target sentences were appropriate in context, and listeners' ratings may reflect that. Of course, if they were *only* rating the text, we would expect the same high MOS across all stimuli, which was not the case: the speech did also matter. A future experiment could manipulate semantic or syntactic mismatch between context and target.

When evaluating speech in context, however, we are most interested in whether the MOS paradigm is sensitive to prosodic differences between stimuli. When modelling context, we expect that an utterance generated using contextual features should sound better than a model that does not take context into account. In experiment 3, we therefore tested whether MOS is sufficiently sensitive to measure differences in prosodic realisation. Clark et al. (2019) showed that varying the contexts between natural speech, synthetic speech and just text led to changes in MOS rating. They postulated that this was due to quality mismatches, with a natural speech context lowering the perceived quality of the following synthetic target. Our experiments exclusively used synthetic speech and did not vary the context utterance, so we can rule out any effects caused by differing contexts. We found that participants rated prosodically non-canonical targets as significantly less natural in isolation than canonical targets: so MOS was sensitive to the differences between the stimuli in isolation. Our stimuli generally had substantial prosodic differences (the non-canonical renditions were very different to the canonical ones), so we are unable to say whether MOS would be sensitive to more subtle differences.

For the in-context evaluation, unexpectedly, *both* non-canonical and canonical target utterance received significantly higher ratings for appropriateness when presented in context than the identical utterances presented in isolation and rated for naturalness. Sometimes, a non-canonical form may indeed sound unnatural if heard in isolation, unless a very specific context is provided in which it sounds felicitous. Our stimuli, however, were constructed to ensure that the non-canonical renditions were *infelicitous* to their contexts, which is why we did not expect ratings of appropriateness to still be higher. Though we did find a significant difference between both renditions in context, the average ratings of the renditions were far closer together than in isolation. In Chapter 2, we gave an overview of perception of prosody in context and a possible reason for the increase in ratings for the non-canonical renditions is *expectation effects*.

Expectation effects arise when the linguistic context primes a listener to expect a certain prosodic form given the discourse structure, which in turn can affect how a prosodic form

is perceived. Various studies have shown this to be the case for prosodic prominence perception within single utterances due to syntactic cues (Calhoun et al., 2021) and word frequency information (Cole et al., 2010). Further, the prior utterance can also lead to priming of certain prosodic prominence in a target utterance (Bishop, 2012; Brown et al., 2015; Im et al., 2023, for example). Therefore we can conceive that if listeners are primed to expect prominence on particular words, this expectation could mediate their perception of the acoustic signal to some degree when a non-canonical realisation is presented. Additionally, because listeners are faced with prosodic variability between and within speakers, they can quickly adapt to speakers who use unreliable variation in pitch accent placement (Roettger and Rimland, 2020). This could mean that the presence of context can constrain the interpretation of the target utterance allowing the meaning of the target utterance to be inferred from context alone (especially for the read speech used in this experiment), thus making differing prosodic realisations more acceptable for listeners or allowing them to quickly adapt to individual differences in prosodic realisation. The role of context in guiding listener adaptation to a speaker's prosodic idiosyncrasies could be tested by consistently presenting listeners with non-canonical prosody by a specific speaker in very specific information structural positions and comparing this to a speaker who uses prosody in a non-systematic way in the same information structural positions. If the context allows listeners to completely overlook differences in prosody, then the non-canonical form should be rated similarly across conditions (whether the speaker is consistent in their use of that form or not). Conversely, if participants are using context to adapt to new speakers but would otherwise find the non-canonical form unacceptable in that position, then we should see differences between both conditions because in one condition the speaker is consistent, so there is a learnable pattern and in the other we cannot make stable predictions, so the use of the non-canonical form may be more harshly rated.

The results from experiment 3 and the role that expectancy effects or speaker adaptation might play, calls into question whether the MOS paradigm would be sufficiently sensitive to detect subtle differences between stimuli presented in context, especially in a real evaluation, with potentially more prosodically similar stimuli. In more recent work on context modelling in TTS, the scores between context conditions have been extremely close (Hu et al., 2022; Xue et al., 2022; Xin et al., 2023)). We could therefore be in danger of false positive or false negative results. When comparing a context-aware model vs context-unaware baseline, with renditions which are perceptually distinct in isolation but which receive a non-significant result in context, we might assume that context does not play a role. Conversely, had the stimuli actually been quite similar in isolation, we would not be able to say with certainty whether similar ratings in context were due to the context MOS boost, or not. In experimental paradigms, we often counter-balance the experimental design using a Latin-square design ensuring that the same text is not heard by any individual listener, from multiple systems. A future line of inquiry would be to compare the canonical and non-canonical renditions side-by-side in a preference test. CMOS preference tests have been recently shown to be more effective in comparing multiple systems (Camp et al., 2023).

Finally, if we holistically evaluate MOS for its ability to evaluate speech in context, we can see that rating based on ill-defined terms such as *appropriateness* cannot tell us what aspects of context have been learned by the model, cannot diagnose issues in stimuli (Gutierrez et al., 2021) and cannot tell us how to specifically improve our model. For example, in experiment 3, the non-canonical realisations showed abnormal prominence

and prosodic boundary placement, but MOS scores cannot detect this. Gutierrez et al. (2021) proposed a diagnostic evaluation based on the Rapid Transcription Paradigm, which can allow listeners to identify problematic areas of a stimulus. Other newer work has looked to using multiple features in evaluation beyond naturalness or overall quality (Seebauer et al., 2023).

When evaluating context-aware models, we need to create evaluation paradigms and test set material that can demonstrate whether the model has learned particular context effects. This is not a trivial matter, because context is a broad and ill-defined term (House, 2007). This is especially the case for conversational speech, the topic of this thesis, in which found data is used. When using found data, we will be faced with different dialects, speaking partners, pragmatic effects and a multitude of unobservable contextual factors (Hodari, 2022). This means that there are more factors than just the previous linguistic context to take into account. What is crucial, and what we will see in the case studies in Chapters 6 and 7, is that we ideally want to have a clear idea of what context we are manipulating, and in turn we need to create a targeted evaluation.

Given the results from this study, we will implement the following in our evaluations for the remainder of this thesis:

- We use MOS as an additional metric to estimate general speech quality. For example in Chapter 3, we use MOS as a complementary evaluation to a more targeted side-by-side evaluation. In Chapter 6, we use MOS to choose speakers who have the highest general quality to use in a further targeted evaluation.

- Where MOS is used, we use ordinal mixed-effects models (Christensen, 2022) to account for the lack of independence in the rating data due to listeners making multiple ratings and due to stimuli being rated multiple times.

- Due to the lack of informativity of MOS ratings, when exploring contextual features in conversational speech, in the case studies presented in Chapter 6 and Chapter 7, we use targeted evaluation that assesses specific contextual features.

## 3.7 Conclusion

We replicated the most interesting finding in Clark et al. (2019): that synthetic speech is rated more highly in context. We investigated the source of this effect, considering the instructions to listeners, textual context-dependence, and prosodic felicity. We found that the wording of instructions had a significant effect on the final MOS score. Instructions that asked listeners to rate *naturalness* resulted in the same rating regardless of whether utterances were presented in isolation or in context. In contrast, asking listeners to rate *appropriateness* of utterances presented in context resulted in a rating higher than the naturalness score, as in Clark et al. (2019). Naturalness and appropriateness are fundamentally different things. It is important, when reporting listening test results, to also report the exact wording of instructions to listeners.

To understand how listeners are interpreting appropriateness, we manipulated the target sentence text. We found no significant difference in the ratings of context-dependent and context-independent text. This does not mean that text plays no role in

appropriateness rating. Future research could manipulate semantic and syntactic factors to gain a better understanding.

We investigated whether MOS is sensitive to prosody, which will be the main difference between the output of a context-aware model and a context-independent one. We found that, for utterances presented in isolation, participants exhibited a greater preference for canonical renditions, a preference that was maintained for utterances presented in context. MOS can be an appropriate paradigm for evaluating prosodic differences, but because the differences between our stimuli were manipulated to be quite distinct, we caution against using this paradigm for in-context evaluation. This increase in MOS was also found for non-canonical items, although they were constructed to be less felicitous in context. It is therefore still unclear what is exactly taken into account in the appropriateness rating. This work could be extended by including other variations in the instructions to participants, such as attempting to focus their attention on prosody or other specific aspects of context.

# Part III

# Speech Synthesis Methods

# 4

# Using Conversational Found Data to Improve Speech Synthesis Prosody

This chapter is based on the following paper:

**O'Mahony, J.**, Lai, C., & King, S. (2022). Combining conversational speech with read speech to improve prosody in Text-to-Speech synthesis. *Proc. Interspeech 2022* (pp. 3388-3392) doi:10.21437/Interspeech.2022-10167

Author Contributions: The work in this chapter was carried out solely by the first author. The contributions of the second and third author were supervisory in nature and both supervisors contributed to editing and feedback on the published paper which shares significant textual overlap with this chapter.

## 4.1 Introduction

As we saw in Chapter 2, end-to-end speech synthesis models have led to significant quality improvements when synthesising isolated sentences using a model trained on *read* speech. However, these models don't generalise well to unseen styles of genres (Li et al., 2021), such as when training on read speech and synthesising conversational speech (Székely et al., 2019b). As we saw in Chapter 1, the goal of this thesis is to begin to explore conversational speech in context and one of the biggest challenges when synthesising conversational speech is the lack of high quality conversational data which can be used to train speech synthesis models.

Speech synthesis datasets normally comprise of professionally recorded data, usually spoken by a voice talent and recorded in optimal conditions (Baljekar and Black, 2016; Cooper et al., 2017; Dall et al., 2016b). Often, these datasets comprise of isolated written sentences, which have been carefully selected to balance certain features of interest, such as

phonological contexts (Campbell, 2005; Baljekar and Black, 2016). Further, these sentences are usually spoken out of a communicative context (Campbell, 2005). Because of this, they may not be suitable for use in applications such as dialogue systems (Székely et al., 2019b).

More recently, researchers have been looking to *found data* to train TTS models. As mentioned (see section 2.3.4), found data is data which has not been explicitly recorded for the purposes of training speech synthesis models (Saeki et al., 2024). This data can therefore come from many sources, such as podcasts and audiobooks (Baljekar and Black, 2016), and because of this, the quality of the data can vary significantly. The main source of found data that has been heavily adopted in this field is audiobooks, for example the LibriTTS Corpus (Zen et al., 2019) and the LJ Speech dataset (Ito and Johnson, 2017). This source of data, however, is confined to read speech. Read speech corpora do not usually contain spontaneous phenomena, for example, false starts or filled pauses (Székely et al., 2019a). Audiobooks do have an important advantage over using recorded isolated sentences, namely that they can be used for context modelling for read speech applications. However, because they were not elicited in a conversational context (Campbell, 2005), they are not suitable for synthesising contextual conversational speech, though dialogue portions of audiobooks have been used to train conversational style speech synthesis (Piits et al., 2022).

Spontaneous conversational speech can exhibit prosodic and phonetic characteristics that distinguish it from read speech (Beckman, 1997), for example, phonetic reduction (Tucker and Mukai, 2023), faster speaking rate (Andersson, 2013), decreased prosodic range (Hazan and Baker, 2010; Adigwe and Klabbers, 2022) and differing stress placement (Howell and Kadi-Hanifi, 1991). These differences, however, are not always present, and the relationship between features and perceived speaking style is quite complex and can be speaker dependent (Batliner et al., 1995). Furthermore, the features that distinguish these styles are not found in all studies and speakers.

Next to the prosodic and phonetic differences between read speech and conversational speech, conversational utterances are highly context-dependent, exhibiting differences depending on the Dialogue Act, speaker stance, turn-taking position and subtle prosodic changes due to phenomena such as (dis)entrainment. These features are absent in read data (Campbell, 2005). Due to this, synthetic voices used in dialogue systems trained on read speech do not sound conversational and struggle with certain prosodic contours found in conversation which may not be present in traditional speech synthesis corpora, for example, different question types (Adigwe and Klabbers, 2022). Though the large amount of variation in spontaneous speech can pose challenges for TTS modelling, it has been found that TTS voices trained on conversational speech are rated as more natural than when trained on traditional isolated prompts (Dall et al., 2014b). This highlights the importance of moving towards training speech synthesis models on spontaneous speech data.

Unfortunately, publicly or academically available datasets of conversational speech have often been developed for large-scale ASR training; such datasets include Switchboard (Godfrey and Holliman, 1993), Fisher (Cieri et al., 2004) and CallHome (Canavan et al., 1997). In these datasets, it is desirable to have data elicited from a large number of speakers. These datasets therefore often comprise of short conversations (between 10-30 minutes) making it difficult to find a large quantity of conversational data from a single speaker. Further, these datasets all comprise of speech recorded via telephone interactions, thus the

recording quality is not optimal for training speech synthesis models. To generate more natural conversational speech we need to look to new sources of data containing these phenomena (Székely et al., 2019b).

In this chapter, we explore a method of using found data from podcasts in The Spotify 100 000 Podcast Dataset (Clifton et al., 2020). Specifically, instead of recording *new* data, we use a target read speech speaker, and *enrich* the training corpus with spontaneous data to increase the prosodic coverage in the speech synthesis training data. Because speech is inherently heterogeneous, due to the wide range of contexts and functions in which it is used (Beckman, 1997), we narrow our focus to the question-answer adjacency pair (Sacks et al., 1974). We chose this particular pair due to its ubiquity in speech synthesis applications, such as dialogue systems. Finally, though it is only a subset of conversation, the prosodic realisation of both questions and answers is highly dependent on the interactional context in which they are found, so they are a good test case before tackling the wider properties of conversation in general.

In this chapter we answer the following research questions:

**RQ4.1** : Does enriching a corpus of read speech with found podcast data improve the prosodic coverage of the target speaker?

**RQ4.2** : Do we find increased preference and naturalness ratings for the model output when training with both read and spontaneous data?

**RQ4.3** : Do we find differences in the prosodic realisation of speech synthesised with a model trained on read and spontaneous speech compared to a model trained on read speech?

To do this, we:

- create a corpus of question-answer pairs from found two-party spontaneous podcast data;
- create a multi-speaker model using read monologue data combined with the above data;
- evaluate whether adding spontaneous data from many speakers improves prosody of questions and answers using both objective and subjective evaluation.

## 4.2 Related Work

### 4.2.1 Recording New Data

In order to synthesise more conversational-sounding speech, different methods of data collection have been proposed. For example, Zandie et al. (2021) created the RyanSpeech corpus by recording a male speaker reading isolated sentences from chatbot scripts in conversational style. Though this approach should lead to an increase in prosodic coverage and conversational style, the speech is not spontaneous because it is still *read* speech spoken by a single speaker. Transcriptions of spontaneous speech read aloud differ significantly from the actual spontaneous speech from which the transcriptions were taken (Guo et al.,

2021). Further, Zandie et al. (2021) recorded isolated conversational utterances, thus losing prosodic phenomena which arise from interaction with another speaker, and other important contextual information coming from prior turns. Because of this, these data are not suitable for training context-aware models (Guo et al., 2021). A more recent conversational dataset, which was published after the work presented in this chapter, is DailyTalk (Lee et al., 2023). Like RyanSpeech, this dataset is based on reenacted speech from a corpus of written conversations. In this dataset, however, the reenactments take place between two speakers and are therefore recorded in context. The speakers are non-native speakers of English and the reenactments suffer again from lack of truly spontaneous speech behaviour.

To overcome the known deficiencies of read conversational style corpora, Guo et al. (2021) used a more natural, yet controlled, approach during corpus creation, recording 45 conversations between two female speakers with semi-scripted interactive scenarios. The speakers were allowed to deviate from the script, permitting phenomena such as false starts. This form of recording also ensures that each utterance is embedded in a communicative context. The conversations were then manually transcribed and used to train a context-aware speech synthesis model.

Adigwe and Klabbers (2022) similarly recorded a conversational dataset, but in three different recording formats. The goal was to investigate how the choice between recording isolated dialogue sentences, recording reenacted dialogue with two speakers or recording a semi-spontaneous dialogue affects the resulting characteristics of speech. They found significant differences between all recording formats. This highlights the fact that reenacted dialogue and reading isolated dialogue utterances do not result in the same speech characteristics as spontaneous speech. They found that read prompts had a higher mean $F_0$ than scripted or semi-spontaneous speech and that the $F_0$ range was higher for read speech, though they interestingly found that there was more variation of $F_0$ range between utterances in the semi-spontaneous dialogue. This, they hypothesised, was likely due to the larger amount of variation and expressivity in semi-spontaneous dialogue. They concluded that while reenactments of scripted conversation lead to prosodic characteristics which are closer to semi-spontaneous speech, speech from spontaneous conversations would still be more suitable for embodied interaction (Adigwe and Klabbers, 2022).

### 4.2.2 Using Found Data

The approaches mentioned above involve recording *new* data, however this is time-consuming and expensive. Other work has taken a different approach by using existing sources of conversational data in the form of *found data*. For example, Cooper et al. (2017) used an ASR dataset of conversational speech used in the creation of dialogue systems to train HMM speech synthesis models. The goal here was not specifically to synthesise conversational speech, but to use English found data to mimic the use case of using such data in a low-resource setting. These data are comprised telephone conversations and as mentioned suffer from poor audio quality. To ensure the model produced intelligible speech, they labelled the data based on a number of acoustic features, such as mean $F_0$, standard deviation of $F_0$, speaking rate etc. They also filtered out data which comprised short utterances, data containing noise based on the transcriptions etc. They found the highest intelligibility was achieved, as measured using word error rate, when they removed data with transcribed noise, reducing the WER from 67.7 to 58.9.

With regards to the acoustic features, they found that high speech rate, low articulation level and high mean energy and standard deviation of energy resulted in the most intelligible speech (Cooper et al., 2017). Further, they created smaller subsets of filtered data, and found that training on cleaner data, even if the data quantity is heavily reduced, can lead to better voices (Cooper et al., 2017). This work shows the importance of data filtering when using found data.

Székely et al. (2019b) looked to podcasts as a source of spontaneous speech. They sourced their data from a podcast series with two hosts, therefore comprising conversational spontaneous speech. Unlike telephone speech, podcasts are often recorded in a studio setting and therefore do not suffer from issues causes by low bandwidth. However, podcast data is most often single channel and therefore if overlapping speech is found it cannot be disentangled. To mitigate this, after segmenting the data using breath groups, the utterances were manually checked for noise, laughter and overlaps (Székely et al., 2019b). They first tested an array of training strategies, including using graphemes versus phones as input, and transfer learning from a read speech corpus versus no transfer learning. They found that the model using transfer learning with phone input produced the fewest number of pronunciation errors. Specifically, they synthesised 400 sentences with each system variant and found 13 pronunciation errors using phonemic input and fine-tuning compared to 49 pronunciation errors for the graphemic input and fine-tuning system and 43 errors for the system trained solely on podcast data with phonemic input. They then compared synthetic speech generated from models trained on read speech, lab-recorded spontaneous speech, and found podcast speech, and evaluated which voice was most appropriate in different genres, for example, reading audiobooks and conversation. They found that voices trained on the podcast data were judged by listeners to be more *appropriate* for casual conversations and public speaking (Székely et al., 2019b).

### 4.2.3 Datamixing Approaches

The work by Székely et al. (2019b) shows the potential of podcasts as a source of truly spontaneous speech which leads to higher ratings of appropriateness for certain genres over read speech data. However, in common with all the other approaches described so far, this approach still involves using a new speaker. This means that the resulting synthetic voices have different speaker identities, but real use cases may demand a single speaker identity for both read and conversational speaking styles. So, we follow Székely et al. (2019b) by using podcast data but, in contrast, we use this naturally-produced spontaneous speech to enrich the prosodic repertoire of a target speaker based on read speech by taking a *datamixing* approach.

Datamixing strategies have been used in previous work modelling spontaneous speech phenomena. For example, Andersson (2013) created both HMM and unit selection voices using utterances from both spontaneous conversations and read speech prompts. Crucially, both sets of data were from the same speaker. He found that, for unit selection voices, combining both styles of speech led to a significant increase in preference for the combined voice compared to a read speech voice when participants were asked which voice had *the most spontaneous speech quality* (Andersson, 2013, p.88), but found that when evaluating overall quality, the read speech voice was significantly preferred. For the HMM voices, he found that a voice trained on spontaneous data alone led to lower ratings of naturalness, but higher ratings for conversational style. When models were trained on the

combined read and conversational data using a context tag to differentiate the styles, he found that, when comparing utterances synthesised with the conversational tag compared with the read speech tag, the read tagged utterances were rated higher for naturalness and no significant difference was found between how conversational both systems sounded.

Following on from the work in Andersson (2013), Dall et al. (2016b), used various datamixing strategies to improve the synthesis of filled pauses after finding that training on spontaneous speech alone led to read speech being preferred during subjective evaluation. They found that using a data labelling strategy like Andersson (2013), led to the highest ratings in a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA)-style test. Specifically, models trained with data marked as read speech or with data marked as read speech with filled pauses marked as spontaneous, were rated above models trained on data marked as spontaneous, models trained on all data pooled with no labels and models created using adaptation from spontaneous or read speech.

In this work, we also take a datamixing approach, but contrary to the work described above, the conversational data are not available for the read speech speaker. Instead, we use utterances from many different speakers found in podcast data, similar to Székely et al. (2019b). By doing this, we can mitigate the issues with finding data in both styles from a single speaker. Further, by using many different speakers, sometimes with only one utterance per speaker, we can mitigate the need to find a large quantity of data from a single speaker. However, having no conversational speech from our target speaker means that we cannot use the approach of labelling the different styles as conversational or non-conversational used by Andersson (2013) and Dall et al. (2016b) because there is a confound between speaking style and speaker identity.

In this chapter, we use more recent synthesis methods than those used in the previous work on datamixing. Previous work by (Latorre et al., 2019), on data imbalance has shown that training a multi-speaker model using the sequence-to-sequence model Tacotron2 (Shen et al., 2018) can lead to better model stability than a speaker-dependent model, even when less data is available for the target speaker in the multi-speaker model. Interestingly, (Latorre et al., 2019) found that having high diversity of speakers leads to more model stability. Similarly, Luong et al. (2019) found that training a multi-speaker model on imbalanced data from multiple speakers led to a higher percentage of preference than the speaker-dependent models, especially for speakers with the least amount of data in the corpus. This suggests that speakers with a lack of data benefit from co-training with other speakers (Luong et al., 2019). In this chapter, we test this with prosodic patterns to investigate whether training using a multi-speaker model increases the prosodic coverage of a speaker who does not have conversational prosodic patterns in their training data.

## 4.3 Data

### 4.3.1 Read Speech Data

Our target speaker dataset is LJ Speech (Ito and Johnson, 2017) which consists of 13,100 utterances from audiobooks read by a female speaker of American English.

### 4.3.2 Spontaneous Speech Data

The Spotify 100,000 Podcast dataset (Clifton et al., 2020) contains 2 TB of data from a selection of 100k podcasts which have been automatically transcribed, punctuated and speaker diarised using Google Cloud Services. The podcasts are all given in single channel format and therefore contain overlapping speech which cannot be disentangled. The podcasts sometimes contain background music or laughter, and the automatic transcriptions contain errors in word recognition and diarisation. Filled pauses and hesitations such as *uhm* are not reliably transcribed. Since we do not have the resources to manually correct or even to quantify the above errors, we applied several stages of filtering to discard suspect data; this has become an important step for any work on found data (Cooper et al., 2017).

#### Data Filtering

We split the data into subsets according to the number of speakers detected by speaker diarisation and retained only podcasts containing exactly two speakers, to obtain ∼74k podcasts. Based on the automatically-generated punctuation and diarisation, we split the transcripts into utterances. Errors in diarisation or punctuation led to some utterances being attributed to two speakers. We retained only utterances attributed to a single speaker and whose transcript was a complete sentence (according to the automatic punctuation).

To extract question and answer pairs, we located utterances which ended in a question mark, though this might exclude questions without typical question syntax, e.g. declarative questions. To extract the answer, we simply took the following turn if attributed to the other speaker as per the diarisation. The speech corresponding to the extracted question-answer pairs forms our corpus. This resulted in 123,943 question-answer pairs. We removed question-answer pairs containing symbols or numbers to avoid text normalisation issues, as well as recordings under 500 ms or over 15 s in duration. The resulting set at this stage contained 92,478 question-answer pairs.

#### Audio Filtering

To ensure that each question and its answer were actually spoken by different speakers (recall that the provided diarisation is imperfect), we extracted speaker embeddings using ECAPA-TDNN (Desplanques et al., 2020) for both and performed speaker verification. At the time, this model was the state-of-the-art in the task of speaker verification and was readily available via SpeechBrain[1] (Ravanelli et al., 2021). We removed question-answer pairs for which the model deemed the speakers were the same. The audio data is single channel and therefore does not offer the possibility to separate speakers by channel. So we used Pyannote audio[2] (Bredin et al., 2020) to detect overlapping speech and removed pairs in which any overlap was found. Finally, we also used a laughter detector[3] (Gillick et al., 2021) and removed any pairs in which laughter was found. The final set comprised 26,876 question-answer pairs amounting to ∼18 hours of questions and ∼20 hours of answers.

[1] `https://speechbrain.github.io`
[2] `https://github.com/pyannote/pyannote-audio`
[3] `https://github.com/jrgillick/laughter-detection`

## 4.4 Method

Our method involves training speech synthesis models on a combination of spontaneous and read speech.

### 4.4.1 Data Selection

For the current work, we randomly selected an equal number of hours of questions and answers from the question-answer dataset (henceforth simply 'spontaneous speech') described in the previous section. All selected utterances had a duration for 1 s to 10 s and a podcast country label of UK, US, Canada, or general English. Note that these labels do not inform us about any dialectal or accent information, therefore we do not account for accent variation in our model. For a baseline read speech model we randomly selected 20 hours of data from LJ Speech in which maximum utterance duration is 10 s (henceforth 'read speech').

In early experiments, we compared models trained on data comprising 0%, 25%, 50%, or 75% spontaneous speech with 100%, 75%, 50%, or 25% read speech respectively. Informal listening showed that the model trained with 75% spontaneous speech + 25% read speech did not suffer significantly in quality compared to using 100% read speech, and that larger prosodic improvements were observed than with the 25%+75% or 50%+50% models. Table 4.1 summarises the data used in subsequent experiments.

Table 4.1: *Approximate training data for each model*

| | read | spontaneous speech | | |
|---|---|---|---|---|
| model | speech | questions | answers | total |
| **baseline** | 20 hours | | | 20 hours |
| **datamix** | 5 hours | 7.5 hours | 7.5 hours | 20 hours |

### 4.4.2 Model

We used FastPitch 1.1 (Łańcucki, 2021), which is a multi-speaker non-autoregressive model with a transformer encoder-decoder architecture. A detailed description of this model was given in Section 2.3.2. FastPitch employs three variance adapters which predict the average values of $F_0$, intensity and duration per input symbol, which in the case of this model were phones. We trained two models. The first model is the **baseline** trained only on read speech (from LJ Speech). Though the baseline is only trained on a single speaker, we initialised a speaker embedding table of the same size as that in our second model, with only one entry being used. Our second model, **datamix**, combines read speech (from LJ Speech) and spontaneous speech (selected from Spotify podcasts using the procedure in Section 4.4.1) in the ratio specified in Table 4.1. The speaker embedding table has 14849 entries: 1 for the single LJ Speech speaker and the remainder for the speakers in the spontaneous speech data; all speaker codes are used during training. We trained each model for 1k epochs with a batch size of 20 on 3 GPUs.

## 4.5 Subjective Evaluation

### 4.5.1 Preference Tests

We hypothesised that **datamix** would generate more natural- and conversational-sounding speech. This was tested in two preference tests using identical stimuli and differing only in the instructions to listeners. In the first test, listeners were presented with 100 pairs of stimuli (each pair comprising the output of **datamix** and **baseline** for the same text) and asked *Which of the following sounds the most conversational?* Stimulus order was randomised within and across pairs, differently for each listener. In a post-test questionnaire we asked the listeners what they understood by the term 'conversational'. The second listening test was identical, except that it asked a new set of participants *Which of the following do you prefer?*

### 4.5.2 MOS Test

To gauge the overall quality of both models we also performed a MOS test which presented another new set of listeners with 100 stimuli, 50 synthesised questions and 50 synthesised statements (counterbalanced across two listener groups so that no participant heard the same text spoken by both systems). Participants were asked to rate each stimulus on a scale labelled *1–bad*, *2–poor*, *3–fair*, *4–good* and *5–excellent*.

### 4.5.3 Stimuli

We started from 100 questions and 100 answers randomly selected from the question-answer dataset and not used for model training. Based only on the natural speech and its automatic transcription, we manually removed utterances containing fewer than 2 words, more than 15 words, profanity, controversial topics, gross grammatical errors, nonsensical content, false starts, or acronyms (to avoid text normalisation errors). From the remaining utterances, we randomly selected 50 questions and 50 answers for use in all listening tests. All textual material can be found in Appendix A.2.

### 4.5.4 Listeners

We recruited ∼30 listeners for each of the 3 listening tests through Prolific[4] who were US residents, native English speakers with no reported hearing impairments, and balanced for sex. Listeners were removed if they did not complete the test, did not use headphones, or had issues playing the audio samples. No listener was permitted to participate more than once.

### 4.5.5 Statistical Analysis

We used mixed-effects regression models to account for lack of independence in the data due to repeated measures for listeners and stimuli. These sources of variance have been found to be quite significant in speech synthesis evaluation studies and pose problems in evaluating TTS output (Rosenberg and Ramabhadran, 2017).

For the analysis of preference test results we use a binomial mixed-effects model using the logit-link function with random intercepts for listener and stimulus to account for

[4] `https://www.prolific.co`

variance between listeners and stimuli. We used the lme4 package (Bates et al., 2015). In this model, we used no predictors and are therefore testing whether the intercept coefficient is different from the null hypothesis, which is that both models have an equal probability of being chosen. This form of testing is roughly equivalent to the exact binomial test, but now we are able to account for random variation due to listeners and stimuli.

For the MOS analysis, we used a cumulative link mixed-model (CLMM) using the ordinal package (Christensen, 2022).[5] These models have been shown to be more suitable for analysis of ratings, as they account for the ordinal nature of the response, and have already been used in Natural Language Generation evaluation (Howcroft and Rieser, 2021). Again, the inclusion of random intercepts allows us to account for variance caused by listeners and stimuli, e.g. listeners using different levels of the ordinal scale. Equations for each model are given in the next section.

## 4.6 Results

### 4.6.1 Preference Tests

31 listeners completed the first preference test, which asked *Which of the following sounds more conversational?*. We removed 1 listener for not using headphones, 1 for having issues with a number of audio files and 1 for completing the test in less time than the total duration of the audio. This left 28 listeners. For both preference tests we used a binomial mixed-effects model with the logit function using the following formula which tests whether the distribution of model preferences differs from chance:

```
choice ~ 1 + (1|listener) + (1|stimulus)
```

Results are summarised in Figure 4.1. We found a significant intercept for questions ($\beta$=0.47 (0.61 prob), CI=(0.56,0.67), $p < 0.01$) which means that **datamix** was chosen significantly more times than **baseline**. For answers, we found no significant difference between **datamix** and **baseline** ($\beta$=-0.17 (0.46 prob), CI=(0.41,0.51), p=0.09).

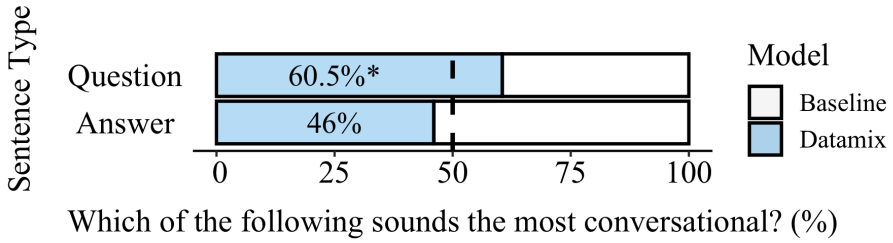


Figure 4.1: *Results for 'Which sounds the most conversational?' Our proposed system is shown in blue.*

The second preference test simply asked *Which of the following do you prefer?* and 32 listeners took part, of which 1 was removed for not using headphones, and 1 for having issues playing

[5]Stimuli and statistical analysis are found here: `https://github.com/johannahom/interspeech22-datamix`

some audio files. The results are summarised in Figure 4.2. Again **datamix** was chosen significantly more times over **baseline** for questions ($\beta$=0.44 (0.61 prob), CI=(0.54,0.67), $p < 0.01$), but not for answers ($\beta$=-0.21 (prob=0.45), CI=(0.39,0.50), p=0.06).

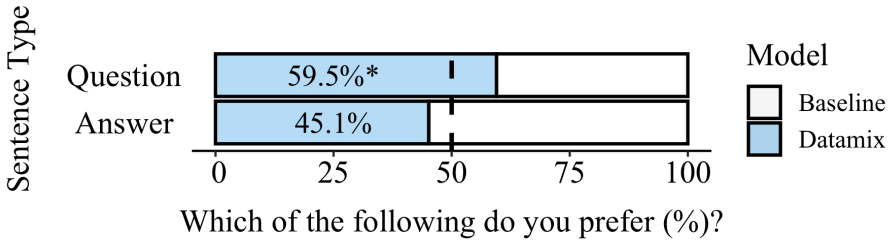


Figure 4.2: *Results for 'Which of the following do you prefer?' Our proposed system is shown in blue.*

### 4.6.2 MOS Test

A total of 62 listeners (30 and 32 per listener group) completed this test, of which 2 were removed for failing to use headphones. The mean MOS for **baseline** when synthesising answers was M=3.19 SD=1.21 and for questions M=2.83 SD=1.20. For **datamix**, the mean MOS for answers was M=3.07 SD=1.24 and for questions M=3.12 SD=1.22. The MOS results are summarised in Figure 4.3 and Figure 4.4.

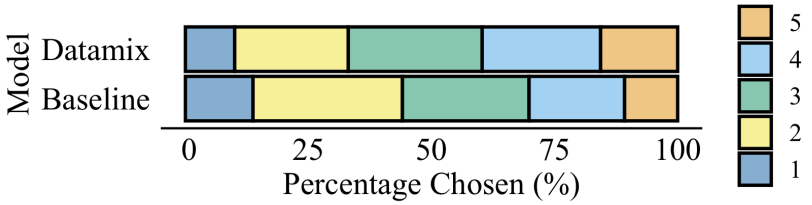


Figure 4.3: *MOS results for questions.*

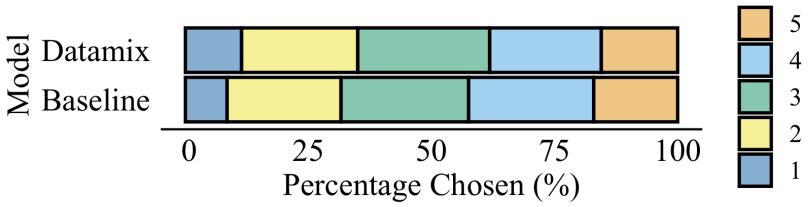


Figure 4.4: *MOS results for answers*

To test whether the models were rated significantly differently, we fitted an ordinal mixed-effects model predicting the effect of each `model` and `sentence type` on the log odds of receiving a particular MOS. We specified a random effects structure to account for repeated measures of both stimulus and listener which accounts for random variation of both, i.e. some listeners will use the scale differently and some stimuli will show

random variation, and a random slope for listeners to account for baseline preference of one model over another. We used the following formula:

```
MOS score ~ model * sentence type + (model|listener) +
                    (1|stimulus)
```

To test the significance of the fixed effect, we compared models using a log-likelihood test between models with and without each factor of interest. We found that there was no main effect of `model` ($\beta$=-0.25, SE=0.25, $G^2$(1)=0.82, p=0.37) or of `sentence type` ($\beta$=-0.74, SE=0.25, $G^2$(1)=3.15, p=0.08). This means that, taken *independently*, there is no significant difference between ratings of questions and answers, or between the models. We did however find a significant *interaction* between `model` and `sentence type` using log-likelihood ratio test between the full CLMM and the CLMM without an interaction factor ($\beta$=0.83, SE=0.35, $G^2$(1)=5.47, p=0.02). To examine this interaction we calculated the predicted probability of each MOS rating per model and sentence type (see 4.5). As we can see, questions in **baseline** have a higher probability than answers of being scored as a 1 or a 2, and consequently a lower probability of getting a higher rating. As in the preference tests, this shows us that **datamix** performs better for questions, but has a lower probability of receiving higher scores than the baseline system for answers.

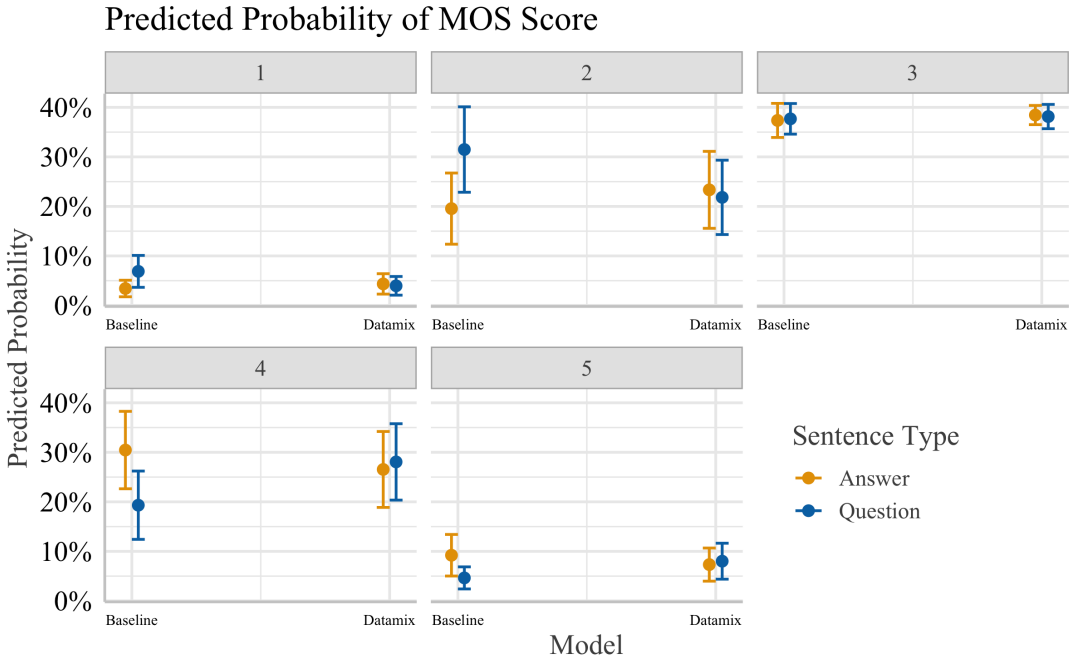


Figure 4.5: *CLMM predicted probabilities of MOS scores for models and sentence types.*

## 4.7 Objective Evaluation

As we saw in the introduction to this chapter, read speech and conversational speech *can* exhibit different phonetic and prosodic phenomena, however this can depend on the specific feature of interest and the speaker. In this study, while we observe differences in preference

between the **datamix** and **baseline** model, it is not clear whether there are acoustic features which significantly differ between the two. This is especially true for the questions which showed a significant difference in preference between the systems. To investigate this, we analysed the prosodic properties of the stimuli generated from the **datamix** and **baseline** model.

### 4.7.1 Feature Extraction

To investigate the differences between the **baseline** and **datamix** model, we first aligned the synthetic speech samples from both systems to phone labels from the CMU dictionary using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). To extract $F_0$, we used the Praat filtered auto-correlation $F_0$ algorithm (Boersma and Weenink, 2024). We used a two-pass method to estimate $F_0$. This involved first extracting the $F_0$ of each utterance with the default pitch floor (50Hz) and pitch ceiling (800 Hz) and after, with the formula described in de Looze and Rauzy (2009), calculating the optimal pitch floor and ceiling values for each utterance. The final $F_0$ contour was then extracted with these optimal values. We removed $F_0$ values in regions marked as a pause or as an unvoiced phone according to the MFA annotations and removed outliers in the global utterance $F_0$ contour if they were above 2.5 standard deviations from the utterance mean.

We then extracted the following $F_0$ features: $F_0$ mean and standard deviation across the whole utterance and the $F_0$ range across the whole utterance using the difference between the 95th and 5th percentile. Similarly, on the final word of each utterance, we analysed the $F_0$ standard deviation, mean and range. To quantify word-final $F_0$ rises and falls, we additionally extracted the $F_0$ slope on the final word. To do this, we removed unvoiced sections of the $F_0$ curve (0 values) and time-normalised the remaining values between -1 and 1. We then fit a linear model using the sklearn LinearRegression function (Pedregosa et al., 2011) and extracted the slope parameter.

Next to the $F_0$ information, we extracted a number of other durational features according to the MFA automatic forced alignments. We first removed leading and trailing silences of each utterance and calculated the total audio duration. Finally, we extracted all word durations per utterance and calculated the mean duration in phones per second.

### 4.7.2 Statistical Testing

All statistical analyses were conducted separately for questions and answers. For questions, the first author labelled each stimulus text as a declarative, disjunctive, polar, or wh-question based on the text of the question and only kept the latter two categories. To investigate whether the speech output of each model differed, we used the Wilcoxon signed rank test (paired two-tailed test). This was used as a non-parametric alternative to a t-test, given that not all of the features followed a normal distribution. Here, the null hypothesis is that there is no difference between each pair of utterances from both systems on each feature of interest.

### 4.7.3 Results for Questions

The results of the comparison of global prosodic features between the **datamix** and **baseline** questions can be found in Table 4.2. Here, we can see that the **datamix** model

had a significantly higher speech rate than the **baseline** model. Features relating to $F_0$ showed no significant differences between systems.

Table 4.2: *Results of comparison between **datamix** and **baseline** model for global prosodic features on questions with a Wilcoxon Ranked Sign Test (significance at $< 0.05$). All values are given in Hertz.*

| | Baseline | | | Datamix | | | |
|---|---|---|---|---|---|---|---|
| Feature | Median | Mean | Std Dev | Median | Mean | Std Dev | Significant |
| Speech Rate (phones/s) | 10.74 | 10.61 | 1.39 | 11.35 | 11.32 | 1.55 | ✓ |
| $F_0$ Mean (Hz) | 215.10 | 218.34 | 15.34 | 219.02 | 222.00 | 17.36 | ✗ |
| $F_0$ Range (Hz) | 98.98 | 103.65 | 26.96 | 100.88 | 102.30 | 28.66 | ✗ |
| $F_0$ Std. Dev (Hz) | 30.87 | 32.12 | 8.24 | 30.25 | 32.26 | 10.03 | ✗ |
| Word Duration (s) | 0.24 | 0.28 | 0.16 | 0.21 | 0.26 | 0.16 | ✓ |

In the analysis of the final word $F_0$, we split questions into polar and wh-questions for the analysis. This is because the canonical forms of these question types typically show differences in their $F_0$ patterns (though this is not always the case when looking at realistic language production, and across dialects see Grabe et al. (2005)). As we can see from the results in Table 4.3, the systems had no significant differences on the features of interest for the polar questions. In particular, when we look at the word-final slope, both systems have a positive median slope. When we look to the wh-questions, however, which canonically end in a fall, we see that the **baseline** system produces more rises than falls, while the *datamix* model produces more falls, however this difference is not significant. We also observe significant differences in $F_0$ range and standard deviation on final-words. In Figure A.1 found in the appendix, we observe that the **baseline** model actually produces a median higher slope on final words in wh-questions, which may have been dispreferred by listeners. Although the characteristic rises in polar questions and falls in wh-questions are not necessarily always found, especially in particular dialogue contexts, it is possible that in this evaluation, where utterances are presented only in isolation, speakers prefer prototypical realisations.

Table 4.3: *Results of comparison between **datamix** and **baseline** model for final-word prosodic features on questions with a Wilcoxon Ranked Sign Test (significance at $< 0.05$)*

| | | Baseline | | | Datamix | | | |
|---|---|---|---|---|---|---|---|---|
| | Feature | Median | Mean | Std Dev | Median | Mean | Std Dev | Significant |
| POLAR | $F_0$ Mean (Hz) | 206.56 | 216.92 | 36.81 | 220.32 | 220.29 | 35.51 | ✗ |
| | $F_0$ Range (Hz) | 73.90 | 70.95 | 31.44 | 68.37 | 59.76 | 34.54 | ✗ |
| | $F_0$ Std. Dev (Hz) | 24.94 | 24.32 | 11.64 | 24.00 | 20.28 | 11.54 | ✗ |
| | Slope | 1.77 | -9.67 | 28.28 | 3.95 | 1.10 | 28.90 | ✗ |
| WH-Q | $F_0$ Mean (Hz) | 208.97 | 209.16 | 20.38 | 203.14 | 207.52 | 25.13 | ✗ |
| | $F_0$ Range (Hz) | 66.74 | 71.93 | 32.02 | 43.05 | 48.08 | 24.35 | ✓ |
| | $F_0$ Std. Dev (Hz) | 22.74 | 23.61 | 9.70 | 15.62 | 17.07 | 8.64 | ✓ |
| | Slope | 6.08 | 3.91 | 30.62 | -10.05 | -7.32 | 21.94 | ✗ |

### 4.7.4 Results for Answers

For the global prosodic features of answers, we see several differences between the output of the models. Unlike the questions, here we see no difference in speech rate or word duration between the **baseline** and **datamix** model. For $F_0$, we see that answers from the **datamix** model tend to have a higher $F_0$ mean, range and standard deviation.

Looking at the final word, we see that both systems end in a median falling slope which is an expected pattern for statements spoken canonically. Again, on the final word we observe a higher $F_0$ in the **datamix** model.

Table 4.4: *Results of comparison between **datamix** and **baseline** model for global prosodic features on answers with a Wilcoxon Ranked Sign Test (significance at $< 0.05$)*

| | Baseline | | | Datamix | | | |
|---|---|---|---|---|---|---|---|
| Prosodic Feature | Median | Mean | Std Dev | Median | Mean | Std Dev | Significant |
| Speech Rate (phone/s) | 9.81 | 9.82 | 1.41 | 9.97 | 9.81 | 1.71 | ✗ |
| $F_0$ Mean (Hz) | 200.93 | 201.98 | 13.27 | 224.82 | 230.11 | 23.31 | ✓ |
| $F_0$ Range (Hz) | 90.52 | 92.24 | 29.79 | 119.73 | 120.93 | 30.78 | ✓ |
| $F_0$ Std. Dev (Hz) | 27.75 | 28.85 | 9.30 | 37.07 | 38.82 | 11.20 | ✓ |
| Word Duration (s) | 0.27 | 0.30 | 0.27 | 0.27 | 0.30 | 0.27 | ✗ |

Table 4.5: *Results of comparison between **datamix** and **baseline** model for final-word prosodic features on answers with a Wilcoxon Ranked Sign Test (significance at $< 0.05$)*

| | Baseline | | | Datamix | | | |
|---|---|---|---|---|---|---|---|
| Prosodic Feature | Median | Mean | Std Dev | Median | Mean | Std Dev | Significant |
| $F_0$ Mean (Hz) | 183.62 | 184.78 | 21.12 | 200.81 | 206.37 | 30.35 | ✓ |
| $F_0$ Range (Hz) | 58.86 | 59.52 | 24.50 | 53.94 | 63.96 | 42.89 | ✗ |
| $F_0$ Std. Dev (Hz) | 19.38 | 19.80 | 7.81 | 17.23 | 22.31 | 16.06 | ✗ |
| Slope | -15.42 | -16.39 | 17.10 | -10.24 | -16.18 | 32.39 | ✗ |

## 4.8 Discussion

In this chapter, we investigated whether training on read and spontaneous speech could bring benefit to a read speech target speaker, especially on categories of dialogue acts, such as questions, which are are likely to be underrepresented in a read speech corpus. The preference test results show that enriching training data with spontaneous speech (in this case from multiple speakers) leads to an increase in listener preference and ratings when synthesising questions. There may be a trend towards a lower preference for **datamix** answers, but this is not significant. The MOS test paints a similar picture: questions were rated significantly higher for **datamix**, but for answers we see a trend of lower ratings of answers from the **datamix** model. This suggests, similar to the work on multi-speaker training with sparse data, that training on many speakers does not significantly hinder the quality of the model. Additionally, adding unrepresented dialogue acts, such as various

types of questions, from many different speakers does bring additional benefit to the target speaker. Thus, despite the fact that we conditioned on a speaker label, we still see some influence of features learned across speakers. Though in this chapter we focused on questions, we could apply this datamixing strategy to a wider number of prosodic constructions which are also not abundant in read speech, such as backchannels, or similar to Andersson (2013) and Dall et al. (2016b), filled pauses and discourse markers.

Comparing the results of the *Which sounds the most conversational?* preference test with those of *Which do you prefer?*, we see roughly similar patterns and we could conclude that listeners did not find our models significantly more *conversational* because the results of both preference tests are quite similar. However, there are alternative explanations for the similar results in both preference tests. As mentioned in Section 4.1, Dall et al. (2014b) found that the output of a TTS model trained on conversational speech is rated as more natural than the output of a model trained on read speech, when the term *naturalness* is not specifically defined. A possible explanation for the similarity in results across experiments is that the textual content is extracted from real conversations and thus these tests are actually asking the same thing. In other words, we would expect the preferred rendition to be the rendition that sounds more *conversational* to match the conversational textual content. At the same time, the slight reduction in preference for answers suggests that the model didn't overall become more conversational. If this were the case, we would expect the model to also perform better for these cases.

One explanation for the lower ratings for answers in the **datamix** model is that the read speech data already contains a large amount of declarative utterances. Because of this, it is conceivable that most of the benefit ultimately comes by enriching the data with a dialogue act class which is heavily underrepresented in the read speech data. We also expect, given the fact that we used speaker conditioning, that the speaker label maintains most of the speaking characteristics of the target speaker. Though speaker and style leakage do happen in multi-speaker models (Stan and O'Mahony, 2023), overall this training scheme is expected to maintain the identity of the speaker, which might include the prosodic phenomena that are tied to their speaking style.

An additional explanation for the lower ratings of answers is that there is a key difference in the selection of the questions and answers in our corpus. Recall that each of the questions and answers was originally based on a question-answer pair. During data filtering, we ensured to the best our ability that the question and answer came from two distinct speakers in the conversation, firstly by using the speaker tags from the original diarisation in the corpus and then secondly validated by the use of the external speaker verification model. This means that the question was always a turn-final construction of the first speaker, but the answers could potentially be turn-medial or turn-final for the second speaker because we do not know who spoke afterwards. We could therefore conceive that answer $F_0$ contours contain a lot more variation than the read speech declaratives, for example ending in final rises or falls due to turn position, pragmatic reasons, or phenomena like uptalk (see Ben-David and Shechtman (2021)). Possibly, answers presented out of context in a listening test will be preferred if they have a prototypical declarative contour and indeed in the objective evaluation we saw a trend that the answers in the **baseline** model ended in a steeper fall. In Chapter 7, we therefore further investigate the impact of turn position on the resulting speech synthesis.

## 4.9 Future Work

There are some improvements to the method in this chapter which could be explored in future work. First, in this chapter, we focused on filtering methods which mainly addressed some of the known issues with automatically transcribing and diarising audio from a single channel. We therefore focused on verifying speaker identity, identifying overlapping speech and annotating laughter. However, we also referenced prior work in the background section which filtered data based on prosodic and phonetic characteristics. Future work should focus on filtering and quantifying the prosodic characteristics of the training data, to allow for increased prosodic coverage or to filter data with undesirable characteristics. Future work could also focus on increasing the quality of the speech output by creating cleaner subsets of data. For example, since the publication of this chapter, newer models have been developed to automatically label audio characteristics, such as speech-to-noise ratio (SNR) (Lavechin et al., 2023) and there have been advancements in speech recognition which allow for better transcription of the textual content. In Chapter 7, we therefore also utilise these newer models to improve the quality of our data and transcriptions.

Second, while in this thesis we are mostly concerned with the prosody of conversational speech, the use of found data from unknown speakers presents a risk that dialectal variation is not accounted for. This variation can be found on the segmental level in the form of pronunciation differences between dialects, as well as the suprasegmental level in the form of systematic prosodic differences between dialects. For example, the found data used in this study comes from many different speakers, and The Spotify 100 000 Podcast Dataset does not provide metadata with information about the speakers involved, such as their dialect or language background. This means that there are potentially many different dialects of English which show considerable differences in pronunciation. Future work should therefore seek to label these differences, manually or automatically, to improve the model by either removing certain dialects, or by accounting for this dialectal variation in the form of model conditioning or using alternative transcriptions.

Furthermore, in Chapter 1, we saw that spontaneous speech can contain more reduced pronunciation variants due to phonetic and syllabic reduction. Again, this has not been accounted for in our model. Unfortunately, most lexica comprise canonical transcriptions and potentially only a select number of reduced variants for commonly occurring reductions. This means that forced alignment will impose canonical segments onto reduced variants, even if not all phones or syllables have been realised. In this work, the FastPitch model uses an internal aligner (Badlani et al., 2022) that is trained simultaneously with the speech synthesis model. Nonetheless, the transcriptions that the aligner is trained on come from the CMU dictionary which to our knowledge does not contain extensive numbers of reduced variants. To account for the deviant realisations in spontaneous speech, future work could explore the use of self-supervised discrete units to transcribe conversational speech data, similar to the method explored in Wells et al. (2023). In this method, a pre-trained self-supervised speech model, such as HuBERT (Hsu et al., 2021), can be use to extract embeddings from an intermediate transformer layer. By clustering the extracted embeddings, discrete clusters of phonetic information can be found and subsequently used as input to a speech synthesis system. This method can be applied in low-resource settings where limited resources, such as lexica, are available (Wells

et al., 2023). Similarly, using this method we could overcome the limitation we have when synthesising conversational speech due to having mismatches between canonical transcriptions and the phonetically realised form. This could potentially allow us to model pronunciation variants that are more commonly found in spontaneous or conversational speech, allowing us to synthesise more conversational-sounding speech.

Finally, in this chapter, we find a clear benefit in enriching a corpus of read speech with spontaneous speech for questions, but overall the speech of the target speaker didn't significantly become more conversational. In the objective evaluation, we did not see consistent changes across questions and answers, which suggests no systematic shift in acoustic characteristics occurred. Again, this is expected, as the speaker label should preserve important characteristics of the target speaker's speech. To achieve more conversational-sounding speech, in the next chapter, we investigate whether we can control the duration and the $F_0$ of a synthesis model to achieve more conversational-sounding speech. Specifically, we introduce a method to control the $F_0$ of a synthesis model hierarchically, on both the word- and phrase-level, and then we in Chapter 6, we explore how we can use found data and controllable synthesis to investigate the prosody-pragmatics interface.

## 4.10 Conclusion

We have shown that enhancing training data with speech from real spontaneous conversations leads to improvements in the ratings of synthetic speech for a target speaker for whom we only have read speech. The introduction of speech from several thousand speakers did not lead to a reduction in quality for the target speaker, and did improve listener ratings of questions. This benefit was not seen for answers. We discussed additional improvements to the method proposed in this chapter, especially with regards to using newer data filtering techniques (to be addressed in Chapter 6) and accounting for turn-position of utterances taken from conversation (to be addressed Chapter 7). Nonetheless, even with minimal data filtering, this method allows us to add prosodic coverage of conversational contours which would otherwise be absent in a read speech corpus.

# 5

# Hierarchical Intonation Control Using Legendre Polynomials

This chapter is based on the following paper:

**O'Mahony, J.**, Corkey, N. Lai, C., Klabbers, E., King, S. (2024) Hierarchical Intonation Modelling for Speech Synthesis using Legendre Polynomial Coefficients *Proc. Speech Prosody 2024* doi: 10.21437/SpeechProsody.2024-208 pp 1030–1034, Leiden The Netherlands

The work presented in this chapter was completed solely by the first author. The contribution of the second author was in creating the initial models on the phrase-level as part of their bachelor thesis and in proofreading the final draft of the above paper. The contributions of the remaining authors were supervisory in nature, including proofreading and editing. Additionally, code was used from Dan Wells relating to the use of ground-truth duration in FastPitch.

## 5.1 Introduction

As we saw in the previous chapter, one of the biggest challenges when synthesising conversational-sounding speech is the lack of suitable data for training speech synthesis models. Many spontaneous speech corpora are not suitable for TTS training, containing either very little data per speaker, or having been recorded in sub-optimal recording conditions. In the previous chapter, we saw that adopting a datamixing strategy, involving the addition of conversational speech to a read speech corpus, can improve the prosody of underrepresented categories, such as questions, in a target speaker for which we have only read speech data. At the same time, we also observed that the improvement was limited to questions and that the method did not systematically change the style of the speech. In this chapter, we therefore introduce a method to control the intonation of a speech synthesis model and in Chapter 6, we use this method to study the prosody-pragmatics interface in

conversational speech.

Controllability in a speech synthesis model is desirable for a number of reasons. First, the ability to control a speech synthesis model allows us to use external prosody prediction models to control the synthesis model at inference time (Watts et al., 2015; Skerry-Ryan et al., 2018; Hodari et al., 2020). This allows us to train the external models with data which may otherwise be unsuitable for training synthetic voices (Ben-David and Shechtman, 2021; Li et al., 2021), or with more data than is needed for voice training, allowing us to capture more variation. In the case of conversational speech, this would allow us to use the conversational corpora that have lower recording quality to learn prosodic patterns using features such as $F_0$, intensity and duration – features which can often be reliably extracted from lower-quality recordings (Raitio et al., 2020). For example, Ben-David and Shechtman (2021) used prosodic information extracted from the Switchboard Corpus of telephone speech (Godfrey and Holliman, 1993) to train a model which predicts prosodic features from text and then subsequently used those features to control a speech synthesis model.

Secondly, as current speech synthesis models reach levels of naturalness close to human recordings on certain read speech datasets (e.g., Tan et al. (2024)), there have been renewed calls for the use of TTS in speech science (Malisz et al., 2019), for example in prosodic research. Typically, in research into the perception of prosody, stimuli are created which differ on certain prosodic dimensions, such as their duration, $F_0$ realisation, or intensity. By creating controllable speech synthesis, we can allow speech scientists to control certain acoustic features to make controlled but realistic stimuli. This form of stimuli creation can be used when certain genres of speech, for example conversational speech, are difficult to elicit in a lab setting. Using TTS to create stimuli has an advantage over traditional methods of stimuli creation in prosodic research, which often involve manipulating an existing recording using signal processing techniques. Such techniques include Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA) for $F_0$ manipulation, splicing, in which words or segments are removed and then reconcatenated with surrounding speech, and duration manipulation. These techniques can lead to signal distortion or unrealistic speech which may harm the external validity of the results.

In this chapter, we focus on the controllability of $F_0$. Some current TTS models have the ability to control $F_0$, on the phone- (Łańcucki, 2021) or frame-level (Ren et al., 2021). However, controlling $F_0$ phone-by-phone, though providing a granular level of control, may not be as useful as control via more interpretable, or theoretically-relevant representations of $F_0$ used in speech science, which is desirable (Grabe et al., 2007). Moreover, to be theoretically-relevant, it is important to model prosodic phenomena on the correct prosodic level (Hodari et al., 2020; Suni et al., 2017), such as the phrase-level or word-level. Other methods of prosodic controllability include using categorical labels, for example using binary labels of prosodic prominence. Such labels encapsulate multiple prosodic features, such as duration, intensity and $F_0$, simultaneously. Next to prosodic features, prominence labels might also encapsulate pronunciation information, for example, the presence of prosodic prominence on words correlates with reduced coarticulation and increased hyperarticulation (de Jong et al., 1993), and differing vowel formant realisation (Mo et al., 2009). While using such labels can provide speech synthesis control, we do not have the ability to disentangle the individual features which are correlated with these labels. Similarly, more recent representations learned via unsupervised learning, such as prosodic style embeddings (Zaïdi et al., 2022) and Global Style Tokens

(Wang et al., 2018) can provide prosodic control of an utterance, but these representations also capture multiple dimensions of prosody and are not guaranteed to capture prosody alone but can capture information about speaker (Zaïdi et al., 2022; Skerry-Ryan et al., 2018) and other non-prosodic information (Raitio et al., 2020). Further, such representations are generally difficult to interpret (Hodari et al., 2020).

In this study, we focus on controlling a speech synthesis model via interpretable $F_0$ representations related to $F_0$ *shape* using the coefficients of Legendre Polynomials. Legendre Polynomials have been used in previous linguistic research to validate prosodic annotations based on symbolic representations used in prosodic research (Grabe et al., 2006, 2007). In this method, the series of polynomials up to a certain order is fit to an $F_0$ contour. The higher the order of the polynomials, the more detailed the contour, however the coefficients of the polynomials are interpretable with regards to the $F_0$ shape only up to the third order. In this chapter, we model the $F_0$ contour hierarchically using a linear regression on each phrase to model phrasal slope, and the first three coefficients of a third-order Legendre polynomial on each prosodically-prominent word to model accent shape. To identify the prosodically-prominent words, we use the Prosody Wavelet Toolkit (Suni et al., 2017), which can be used to extract information about the strength of prosodic prominence and boundaries. The resulting values of acoustic strength can be discretised to create categorical labels for these features. We compare both a Legendre-conditioned model with a categorically-conditioned model to investigate whether adding a sparse representation of $F_0$ shape can increase perceived similarity between the synthesised speech and reference speech.
Specifically, we ask:

**RQ5.1** : Does conditioning a FastPitch (Łańcucki, 2021) TTS model on Legendre coefficients and slope lead to increased similarity to a reference compared to the baseline FastPitch model?

**RQ5.2** : Does conditioning a FastPitch TTS model on Legendre coefficients and slope lead to increased similarity to a reference compared to a FastPitch model conditioned on binary prominence and boundary information?

**RQ5.3** : Does conditioning a FastPitch TTS model on Legendre coefficients and slope lead to increased similarity to a reference compared to a FastPitch model controlled using ground-truth $F_0$ values from the reference utterance?

To do this, we:

1. condition a FastPitch TTS model using phrase-level slope coefficients and word-level Legendre polynomial coefficients on prominent words to provide controllability of the $F_0$ contour

2. condition a comparison model using categorical prominence and boundary markers (Suni et al., 2020);

3. synthesise speech using the baseline model conditioned on ground-truth mean $F_0$ per phone.

4. evaluate prosodic similarity using both objective and subjective evaluation.

## 5.2 Related Work

### 5.2.1 Speech Synthesis Controllability

In Chapter 2, we gave a brief overview of speech synthesis controllability. In short, we can add controllability to a model by providing additional information to the model. If this information can account systematically for variation in the speech, then the model should learn the relationship between the additional information and the corresponding speech. In terms of the overall objective in model training, we are moving from *given this text, predict the mel spectrogram* to *given this text and some other information known about the training samples, predict the mel spectrogram*. At inference, we can generate different renditions of the same utterance by changing the information provided to the model. The form of this additional information is dependent on what is being represented (e.g. which feature we are trying to control) and how this representation is learned or acquired. In this section, we will give a brief overview of some of the methods used to generate prosodic representations which can be used as additional information to condition a model, and we give an overview of how these representations are used in speech synthesis research.

Signal-based methods involve extracting continuous features directly from the speech signal. This can be done by extracting $F_0$ directly using a pitch tracker, or computing other features such as spectral tilt (see Raitio et al. (2020, 2022a,b); Ben-David and Shechtman (2021)). These features can then be used to condition a model. The linguistic level that these features control can vary, operating on the utterance-level, phrase-level, word-level or more granularly at the frame- or phone-level. For example, in FastPitch (Łańcucki, 2021), the model is conditioned on the mean $F_0$ per phone, which provides additional information to the model to aid the prediction of the mel spectrogram, for example by allowing the model to learn the relationship between prosodic features and differences in pronunciation (Łańcucki, 2021) and brings the additional benefit of allowing a user to control this feature per phone at synthesis time (Łańcucki, 2021). Raitio et al. (2020), on the other hand, controls a Tacotron 2 (Wang et al., 2017) at the utterance-level with features which have been aggregated across the utterance, such as $F_0$, and intensity, and spectral tilt. As noted by Raitio et al. (2020), the benefit of using acoustic correlates of prosody derived from the signal is that the representations of each prosodic correlate are already disentangled (though they may be correlated) and that many of these features can be extracted from sub-optimal quality recordings.

Speech synthesis models can also be controlled via discrete labels. For example, prosodic prominence labels have been used to condition TTS models (Suni et al., 2020). Prosodic prominence is a perceptual feature, but it normally correlates with increased $F_0$ modulation, energy, and duration, relative to other words in the utterance. Discrete labels of prosodic prominence can therefore encapsulate multiple features simultaneously. Discrete labels have the benefit of being user-friendly and are often created to be intuitive for users to control, but for the use-case of prosodic research, they do not provide the ability to control each prosodic correlate separately. Secondly, they often require human-labelled data, though, as we saw in Chapter 3 and will see in more detail in Section 5.2.3, separate algorithms have been developed to label data for this feature without the need for human annotators.

Discrete labels can also describe a single disentangled feature. For intonation, for example, the ToBI labelling convention, based on autosegmental-metrical theory, was developed to describe pitch tones and breaks. These labels are based perceptual judgements

(Beckman et al., 2005) of expert annotators, usually in conjunction with a graphical image of the $F_0$ trajectory (Beckman et al., 2005; Grabe et al., 2007). This makes them expensive and time-consuming to create, and means that many corpora with these annotations are probably too small for large-scale speech technology needs (Grabe et al., 2007). Further, despite representing key symbolic events in the speech, these features do not describe the actual phonetic realisation, and the authors note that phonetic details can be described with continuous features (Beckman et al., 2005).

Next to discrete methods introduced above, we can use multi-dimensional prosodic embeddings, which can be learned in an unsupervised manner by constructing a latent space using a *reference encoder* (e.g. (Skerry-Ryan et al., 2018)). Here, the speech signal, typically represented by its mel spectrogram, is fed to a encoder module which outputs a fixed-length vector representation of the speech. The choice of the dimensionality of the resulting representation will naturally affect how much information the representation can capture. This representation can then be concatenated with or summed to the speech synthesis encoder output. Though this method was found to increase the similarity between the synthesised speech and the reference speech, in cross-speaker prosody transfer prosodic aspects tied to the original speaker were also entangled in the representation (Skerry-Ryan et al., 2018).

Extensions to learning unsupervised reference embeddings include Global Style Tokens (GST), which learn a set of style labels from the resulting reference embeddings (Wang et al., 2018), as well as reference embeddings which operate on hierarchical linguistic levels (An et al., 2019), rather than as a global embedding. One of the main issues when using these representations, apart from the potential entanglement of speaker information and prosodic information, is that they are difficult to interpret. More recently, control methods operating on these embedding spaces have been developed, to identify prosodic correlates in the embeddings (Šimko et al., 2023), but again these controls operate on the global utterance characteristics. This makes meaningful control of such models for prosodic research difficult.

To allow for the benefits of discrete labels, without the need for expert annotation, Hodari et al. (2020) proposed using a variational autoencoder (VAE) to learn discrete representations of intonation on the phrase-level. A VAE is an autoencoder framework that imposes constraints on the distributional properties of the latent space learned by the network, using a prior. By structuring the space in particular way, it can be sampled from in a meaningful way (i.e. points closer together share similar properties). This method learns an underlying latent space, but at the same time learns to structure the latent space in such a way that clusters emerge (Hodari et al., 2020). Hodari et al. (2020) used a particular prior which is a Gaussian mixture model (GMM), which is trained with the encoder network. By specifying a number of mixture components for the GMM, discrete intonation codes can be learned by taking the mean of each of the Gaussian components. Hodari et al. (2020), compared an autoencoder architecture, with an additional k-means clustering step on the embeddings, with the VAE architecture. They tested which architecture was able to produce the most distinct clusters of intonational renditions. They found that both systems were able to produce perceptual differences, with the VAE producing more distinct renditions.

As we can see, there are numerous ways to represent prosodic information and these representations can be used to control speech synthesis models. Each form of representation has advantages and disadvantages. In this work, however, our aim is to

provide a disentangled representation of intonation which is of use to speech scientists and interpretable. We therefore take a signal-based approach by extracting the $F_0$ contour from our training samples, and subsequently modelling the shape of the $F_0$ contour on both on the phrase and word-level.

### 5.2.2 $F_0$ Representation: Legendre Series of Polynomials

In this chapter, we focus on controlling $F_0$ and are thus controlling intonation. Intonational modelling has a long history in both speech science and speech technology, and because of this there is a broad taxonomy of approaches to modelling this phenomenon (Reichel, 2011). According to Reichel (2011), models of intonation can be distinguished across three features, namely their "*units* (tones or contours)", their "*description* (symbolic vs. parametric)" and their "*arrangement* (single-layered vs superpositional)" (Reichel, 2011, p.341). The choice of approach is often dependent on an underlying assumption about the nature of intonation (Reichel, 2011).

In addition to the model taxonomy outlined by Reichel (2011), according to Taylor (2000), a good model of intonation should specify enough information to synthesise a wide variety of $F_0$ contours, and specifically, the model should be non-redundant, exhaustive and be "linguistically meaningful'" (Taylor, 2000, p.3). Many models that have been developed to be both linguistically-motivated and of use to speech technologists have not been widely adopted in *recent* speech technology applications. The models which have been used in both fields tend to have been developed before the current state-of-the-art TTS models, such as sequence-to-sequence models. This is because the historically strong relationship between both fields has weakened over time. We therefore take inspiration from previously developed intonational models, in particular their linguistic relevance and interpretability, but we use the representations to condition an E2E synthesis model. In this section, we will give a brief overview of some of the previous approaches to intonation modelling, and we will introduce the approach that we will take in this chapter.

In the Section 5.2.1, we saw an example of a symbolic intonation model, the ToBI model of intonation (Beckman et al., 2005). In this work, however, we are focusing on a parametric representation which characterises the shape of the $F_0$ contour, which requires no expert annotation. Previous work on intonation contour modelling includes sequential models, such as the Parametric representation of Intonation Events model (PaIntE) model (Möhler and Conkie, 1998), which models the $F_0$ contour shape across a three-syllable span using six parameters describing two sigmoid functions. This model has been applied to both TTS intonation modelling and prosodic research (Schweitzer et al., 2022). In addition to sequential models, hierarchical models have been developed which assume multiple levels of intonational structure, usually modelling the prosodic phrase and the pitch accent. Such models include the CoPaSul model (Reichel, 2011), which was developed for modelling and linguistically-analysing the $F_0$ contour, combines a linear phrase component (slope) with word-level third-order polynomial stylisation on pitch accents. It has not been applied to TTS modelling, but has been used to manipulate stimuli using TD-PSOLA.

In this chapter, we take a very similar approach to the CoPaSul model, but we do apply our method directly to modern TTS. Just as in the CoPaSul model, we are taking a superpositional approach by modelling a phrase-level feature and a word-level feature. Specifically, we apply linear regression to the $F_0$ contour of each prosodic phrase and

extract the slope coefficient of the fitted line to represent our phrase-level conditioning information. Mirroring the approach used by Reichel (2011), to model pitch accents, we fit a third-order polynomial to the $F_0$ contour of each prominent word. Specifically, the polynomial coefficients are extracted from the Legendre polynomial series, which we have been chosen due to its previous use in linguistic studies, and due to the orthogonality of polynomials in the series (Grabe et al., 2007).

The Legendre polynomial series of orthogonal polynomials can be used to describe a time-series, such as $F_0$. Each polynomial is multiplied by a specific *coefficient* before the resulting polynomials are summed (Grabe et al., 2006, 2007; Svatošová and Volín, 2023). Crucially, the first three coefficients of the series are interpretable. These coefficients represent $F_0$ height, slope, and convexity respectively (Grabe et al., 2003, 2006; Svatošová and Volín, 2023). The fact that these polynomials are orthogonal means that there is no correlation between each of the polynomials (Grabe et al., 2006), and this helps to ensure our representation is non-redundant, which was an important requirement for an intonation model outlined by Taylor (2000).

Legendre polynomials have been used in previous prosodic research. For example, Grabe et al. (2006) used them to model nuclear accents in various English dialects and compared this representation to annotations based on the autosegmental-metrical framework. They found that Legendre polynomial coefficients could statistically distinguish the accent types analysed in their study. They noted that this representation could be used to automatically detect pitch accents in large corpora, where hand-labelling may not be feasible, and that these labels have the potential to be used in speech synthesis systems (Grabe et al., 2006). In Grabe et al. (2003), Legendre polynomials were used on the utterance-level investigate dialectal differences in English in the realisation of declaratives and interrogatives. They found that the first two Legendre polynomial coefficients were sufficient to account for differences between different question types, and between questions and statements (Grabe et al., 2003). In addition to these studies, Legendre polynomials have been used in other studies of intonation, for example to investigate native versus non-native prosody (Rakov and Rosenberg, 2017; Rakov, 2019), to evaluate mimicked speech (Mary et al., 2013) and to investigate final $F_0$ rises in task-oriented dialogue and conversational speech (Lai, 2014).

Finally, Legendre Polynomials have also been used previously in non-neural TTS, for example HMM synthesis (Hsia et al., 2010; Wu et al., 2010) and to improve Mandarin tone representations in formant synthesis (Lee et al., 1993). More recently, Legendre polynomial coefficients were used to control intonation of prosodic phrases by conditioning a FastPitch model (Corkey et al., 2023). Modelling $F_0$ on just one level, such as the phrase, will only give a coarse specification of an utterance's $F_0$ contour and, as we saw in Grabe et al. (2003), this might effectively characterise differences between questions and statements. But this specification will fail to capture more detailed $F_0$ information, such as word-level pitch accent shape. In our work, we therefore extend this method by using both phrase-level and word-level components. We take the approach of conditioning a model with a hierarchical representation of intonation, rather than using a model to strictly enforce an intonational contour. This means that the model should infer the relationship between the input representation and the resulting mel spectrogram, as mentioned in the previous section. To do this, we need to identify prosodically prominent words and phase boundaries and for this we will use the Continuous Wavelet Transform toolkit.

### 5.2.3 Categorical Features: Continuous Wavelet Transform (CWT)

In this work, we use the Wavelet Prosody Toolkit, which is an implementation of an algorithm that detects prominent words and prosodic phrase boundaries in the signal using the CWT (Suni et al., 2017)[1]. As we saw, categorical representations of prosodic events, such as the presence of prosodic prominence and boundaries, can be used to control speech synthesis models (Suni et al., 2020). Recall that these representations were used in Chapter 3, to control a model and produce stimuli for investigating the evaluation of synthetic speech in context. These categorical representations are a low-dimensional representation of prosodic events (Suni et al., 2017), and thus do not have the granularity of other more explicit acoustic correlates of boundaries and prominence. For example, in the case of pitch accents, they do not represent detailed information about the phonetic realisation of $F_0$, an important feature which can be used to distinguish different pitch accent categories (Grabe et al., 2007, 2006). Nonetheless, they provide important information about where salient prosodic events occur in the speech signal. By identifying salient prosodic events in the speech signal, we can subsequently extract more detailed $F_0$ information on the prominent words and segment utterances into phrases based on the presence of prosodic boundaries.

The CWT algorithm is a completely unsupervised and data-driven algorithm that uses word duration, interpolated $F_0$ values, and interpolated intensity values to estimate prominence and boundaries in the signal. After the signals are combined, either by summing or taking the product, CWT analysis is performed by decomposing the resulting signal into various scales which can be visualised in a scalogram (Suni et al., 2017, p.127). Strength of prominence and boundaries per word can be estimated by finding points of lines of maximum amplitude (prominence) and lines of minimum amplitude (boundaries), as outlined in Section 2.2.1 and described in detail in Suni et al. (2017). The resulting values can be discretised into bins. The number of bins and the threshold at which to discretise the values is at the discretion of the researcher. For prominence, for example, a common number of bins is two, to distinguish prominent from non-prominent words, though an additional third bin can be used for strongly emphasised words.

The CWT method, implemented in the Wavelet Prosody Toolkit, was chosen in this work because to our knowledge it is the best-performing open source model for this task. Previous work shows that it reaches 84.6% accuracy on binary prominence detection for the BURNC corpus and 85.7% accuracy on boundary detection (Suni et al., 2017). Furthermore, the toolkit was chosen as it has been previously used in speech synthesis research (see also Chapter 3). When being used to control a speech synthesis model, the raw prominence and boundary strengths can be used to condition the model (Kakouros et al., 2023), or the discrete labels can be added as mark-up to the the input text (Suni et al., 2020), or the discrete categories can be embedded and summed to the encoder output (Stephenson et al., 2022). (Stephenson et al., 2022) showed that this form of controllability can successfully increase the ratings of prominence on *pronouns* in synthetic speech, but that this effect was not fully consistent (Stephenson et al., 2022). This suggests that this form of controllability will still be subject to constraints imposed by the training data, in other words, it may sometimes be challenging to control prominence on words which are not frequently prominent in the training data.

[1] `https://github.com/asuni/wavelet_prosody_toolkit/`

## 5.3 Method

The goal of the current work is to validate that Legendre polynomial coefficients and slope provide an adequate representation of the $F_0$ contour for use in controllable synthesis. Our method involves performing TTS, but with prosody *transferred* from a reference recording rather predicted from the input phone sequence. We focus on controllability because this model will be used to create stimuli in Chapter 6. In all models, duration is directly copied, per phone, from the reference for testing.

### 5.3.1 Data

We use the LJ Speech corpus[2] (Ito and Johnson, 2017), comprising 13 100 utterances read by a female US English speaker. We use a subset of this data to train all models (see Section 5.3.4)

### 5.3.2 Feature Extraction

#### Prosodic prominence and boundary detection

We first aligned the data with the Montreal Forced Aligner (McAuliffe et al., 2017) to obtain word and phone alignments. To identify prominent words and phrase boundaries, we used the Wavelet Prosody Toolkit (Suni et al., 2017), which was described above in Section 5.2.3. To identify boundaries, we used the sum of the $F_0$, intensity, and word duration signals with weights 1.0, 1.0, and 0.5 respectively. For prominence, we used the product of the signals with weights of 1.0, 0.5, and 1.0 respectively. Discretising the resulting values at a particular threshold results in a categorical prominence or boundary label. Carefully articulated speech, such as read news, has between 52% and 54% prosodically prominent words (Yuan et al., 2005). We therefore used a value of 53% as a heuristic, and discretised the prominence values at the 47th percentile, computed per-utterance. We chose a boundary threshold of 1.0 as this was the minimum boundary value over all utterance-final words. The final result of this step was that for each word in an utterance, there was an associated binary boundary label indicating whether the word was associated with a prosodic boundary or not. Similarly, for prominence, the final result was that each word in an utterance was labelled as either being prominent or non-prominent.

#### $F_0$ Extraction

We used the Praat (Boersma and Weenink, 2021) autocorrelation pitch ($F_0$) estimation algorithm to obtain $F_0$, via the Parselmouth python package (Jadoul et al., 2018). We first estimated $F_0$ with the default parameters: a pitch floor of 75 Hz and a pitch ceiling of 600 Hz. We then calculated the optimum pitch floor and pitch ceiling using the method in de Looze and Rauzy (2009) and re-extracted the $F_0$ with these new values. $F_0$ was transformed to semitones relative to the speaker's $F_0$ median across the entire corpus. We removed $F_0$ values more than 2.5 standard deviations away from the utterance mean $F_0$ to identify pitch-tracking errors.

[2] `https://keithito.com/LJ-Speech-Dataset`

**Slope and Legendre Polynomial Coefficients**

To extract hierarchical phrase- and word-level features, we used the phrase boundaries and prominent words found in Section 5.3.2. We first split each utterance into prosodic phrases by identifying words associated with a prosodic boundary and delimiting the phrase at the end of this word. Our phrase-level feature is the $F_0$ slope. We extracted the slope by performing linear regression on the $F_0$ contour associated with each phrase. Each phrase receives a single slope value.

For our word-level Legendre polynomial features, we first identified the prominent words in the utterance. We then removed the effect of the previously-estimated slope by detrending the $F_0$ values: subtracting the value of the fitted regression line from all individual $F_0$ values in that phrase. We then normalised the times stamps in the words between $[-1, 1]$ (the domain for the Legendre Series), and fitted a third-order Legendre polynomial to the detrended $F_0$ contour associated with each word. Our word-level features comprise the *first* three coefficients of that polynomial, and represent height, slope, and convexity respectively. Non-prominent words and silences received zero values. An example utterance with fitted slopes and word-level Legendre polynomials can be seen in Figure A.2 in the appendix.

### 5.3.3 Models

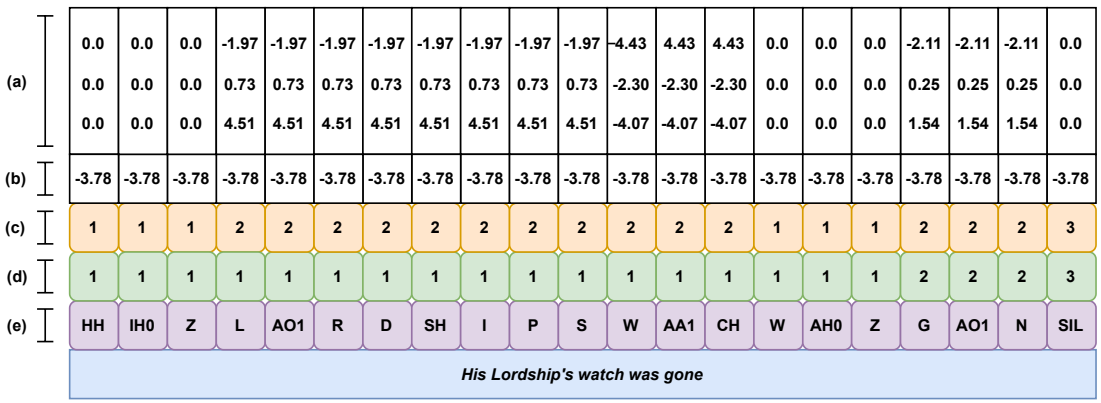


Figure 5.1: *A visualisation of input features of FastPitch. Continuous features: (a) Legendre polynomial coefficients (b) slope; and categorical input features from CWT: c) prominence labels and d) boundary labels*

All of our models are variants of the FastPitch TTS model (Łańcucki, 2021) using our fork of the original code[3]. As we saw in Chapter 2, FastPitch is a non-autoregressive transformer-based model that takes a phone sequence as input and predicts a mel spectrogram. The model contains *variance adapters*, trained to explicitly predict the mean $F_0$ and mean energy of each phone. The model also explicitly predicts the duration of each phone in frames. Unlike the baseline implementation, which uses a neural aligner that is trained along with the speech synthesis model, in our version of the model the duration targets are extracted from the textgrids generated by the MFA aligner. When training the model, ground-truth values for $F_0$ and energy are used both to condition the model's

[3] `https://github.com/johannahom/FastPitches`

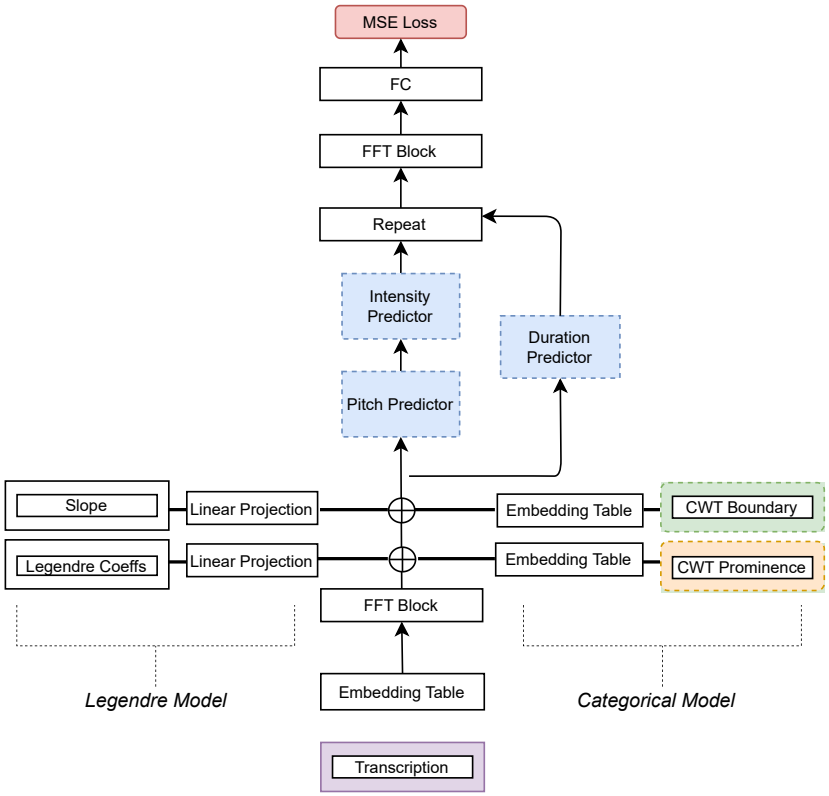


Figure 5.2: *Architecture of FastPitch with continuous or categorical conditioning (adapted from Łańcucki (2021) Figure 1)*

decoder and to train the variance adapters. During inference, either the model's predictions *or* user-provided values can be used, the latter providing a means of control.

### Legendre Coefficient Conditioning

To use externally-provided phrase-level slope and word-level Legendre polynomial coefficients to control $F_0$, we modified the FastPitch architecture: Phrase-level slope values are upsampled to the number of phones in the corresponding phrase within that utterance. This results in one tensor per utterance $S$ with shape $[\text{phone sequence length}, 1]$. An example of these input features can be seen in panel (b) in Figure 5.1. In this example, the utterance comprises a single prosodic phrase, and therefore the slope value is the same throughout the utterance.

The word-level features consist of three Legendre coefficients for each word (the coefficients are set to 0 for non-prominent words). These are upsampled to the number of phones in that word, resulting in a tensor $L$ with shape $[\text{phone sequence length}, 3]$. Again, an example of these input features can be found in Figure 5.1 in panel (a). $S$ and $L$ are each passed through a linear layer to project them to the shape of the encoder output $[\text{phone sequence length}, 384]$ and the resulting tensor is then summed to the encoder output before being passed to the variance adapters, which is shown in Figure 5.2.

**Categorical Prominence Conditioning**

For our categorically-conditioned model, we conditioned a variant of FastPitch model on labels extracted as described in Section 5.3.2. This model receives two externally-provided features. $P$ is a sequence of prominence labels, one for each word in the utterance (1: non-prominent, 2: prominent, 3: silence, 0: padding). $B$ is a sequence of phrase boundary labels, one for each word (1: phrase-internal, 2: prosodic phrase-final, 3: silences, 0: padding). $P$ and $B$ are each upsampled to the number of phones in the corresponding words, resulting in tensors of shape $[\text{phone sequence length}, 1]$. An example of these input features is shown in Figure 5.1 in panel (c) for the prominence labels and (d) for the boundary labels. For each feature, an embedding table was instantiated with an embedding dimension matching the size of the encoder output (384). Both $P$ and $B$ were passed to their respective embedding tables. The resulting embeddings were summed to the encoder output before being passed to the variance adapters, which is shown in Figure 5.2.

### 5.3.4 Training

For the experiments described below, we trained four models. **BASELINE** is an unmodified FastPitch model, which receives no additional conditioning inputs. It predicts $F_0$ using the usual variance adapter. For this model, and all other models below, during inference ground-truth phone durations are used rather than predicted durations. **LEGENDRE** is the model from Section 5.3.3 conditioned on a slope parameter for each prosodic phrase and a set of three Legendre polynomial coefficients for each prominent word. **LEGENDRE** might learn which words are prominent simply by virtue of them receiving non-zero values for polynomial coefficients as opposed to zero values on non-prominent words, and learn prosodic phrase locations from the presence/absence of slope values. To confirm that **LEGENDRE** is actually using slope and coefficient *values* to predict better $F_0$, we introduce the **CATEGORICAL** model. **CATEGORICAL** is the model from Section 5.3.3 conditioned on CWT-derived prominence and boundary labels. Prominent words are those same words that receive Legendre polynomials in **LEGENDRE**. Prosodic phrase boundaries are in the same places that the **LEGENDRE** model receives a new slope value. **ORACLE-GT-F0** is a gold standard, identical to **BASELINE** except that it uses an externally-provided ground-truth mean $F_0$ value per phone (rather than the value predicted by the variance adapter). These $F_0$ values were extracted using a pitch tracker and the mean $F_0$ per phone is calculated. We expect the gold standard will beat **LEGENDRE**, but hypothesise that the difference in listener preference will be small, thus validating that our much sparser and interpretable representation of $F_0$ is adequate.

For each of the models, the same 12691 utterances were used for training[4] with 129 for measuring validation loss. Chapter 50 was reserved as test material, comprising 278 utterances. Each model was trained on a single GeForce GTX 1080 Ti GPU for 500 epochs.

[4] LJ019-0167 and LJ011-0050 removed due to processing issues

## 5.4 Evaluation

Our evaluation method involves prosody transfer from an original naturally-spoken reference recording of the same text as that being synthesised. Because we are only interested in $F_0$, all models perform synthesis using ground-truth phone durations from the reference. Each model (except **BASELINE**) receives externally-provided conditioning, derived from the intonation of the reference recording: **ORACLE-GT-F0** receives the reference recording's $F_0$ value, per phone; **LEGENDRE** receives slopes and polynomial coefficients, fitted to the reference recording's $F_0$ contour; **CATEGORICAL** receives prominence and boundary labels, derived from the reference recording via the CWT.

For both objective and subjective evaluation, 50 utterances were randomly selected from the test material. All generated mel-spectrograms were vocoded using the same pre-trained NeMo Hifi-GAN vocoder[5]

### 5.4.1 Subjective Evaluation

#### Task

We conducted three separate listening experiments to answer our research questions from Section 5.1. The task in each experiment was to listen to the original natural (ground-truth) recording of a test sentence, and make a forced-choice between renditions from a pair of models. We asked listeners *"Which rendition sounds most like the reference rendition?"* The three experiments compared the following pairings of models: **LEGENDRE** vs. **BASELINE**; **LEGENDRE** vs. **ORACLE-GT-F0**; **LEGENDRE** vs. **CATEGORICAL**. Within each experiment, every listener was presented with the same 50 pairs. The presentation order of the 50 pairs, and within-pair order, was randomised per listener. Each listener took part in only one experiment.

#### Participants

Listeners who declared English as their native language and had no known hearing impairments were recruited through Prolific[6]. After the responses were collected, we removed all listeners who did not report using headphones, had difficulty playing audio, or who finished the experiment in less time than the total audio duration, resulting in 16, 19, and 19 listeners for the three experiments respectively.

#### Statistical Analysis

For analysis, we employed binomial mixed-effects regression models with a logit-link function (Bates et al., 2015), without using predictors, which is the mixed-effects equivalence of a binomial test. As we saw in Chapter 4, using a mixed-effects models accounts for the lack of independence of the ratings in the experiment, due to listeners rating the same sets of systems multiple times. We therefore included text (because the text being synthesised will have an effect on the synthetic speech, regardless of model) and listener as random effects.

[5] `https://github.com/NVIDIA/NeMo`
[6] `https://www.prolific.com`

Below is the formula in which *choice* denotes the stimulus that the listener selected as sounding most similar to the reference recording.

```
choice ~ 1 + (1|listener) + (1|text)
```

**Results**

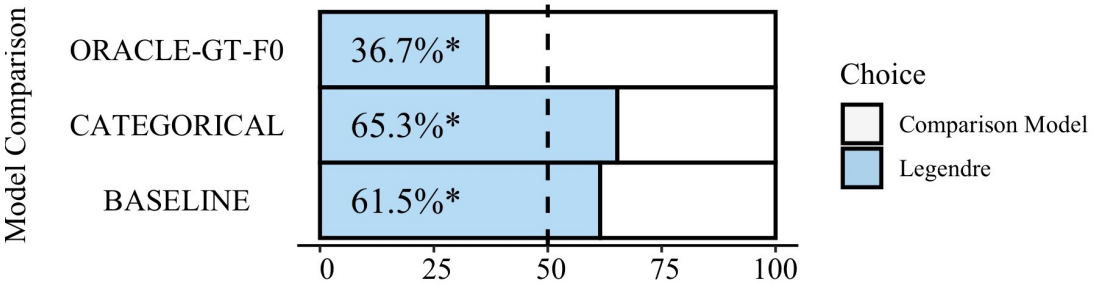


Figure 5.3: *Listeners' pairwise preferences between models*

The results in Figure 5.3 show that our model **LEGENDRE** (in blue) was chosen 61.5% of the time ($\beta$=0.53 (0.63 prob), CI=(0.56,0.70), $p < 0.01$) when compared to **BASELINE**. As expected, listeners preferred the gold standard **ORACLE-GT-F0** more often than our model, with **LEGENDRE** being chosen 36.7% of the time ($\beta$=-0.65 (0.34 prob), CI=(0.27,0.43), $p < 0.01$). Finally, when compared to the **CATEGORICAL** our model was chosen 65.3% of the time ($\beta$=0.68 (0.66 prob), CI=(0.61,0.72), $p < 0.01$). We can conclude that conditioning the model on phrase-level slope and word-level Legendre polynomial coefficients produces more similar-sounding intonation than the baseline and the categorically-conditioned model. As expected, our model does not beat the gold-standard model.

### 5.4.2 Objective Evaluation

In addition to the subjective evaluation above, we also performed two objective evaluations to answer our research questions from Section 5.1. The first of two objective evaluations compared Legendre polynomial coefficients from the original recordings with those recovered from the synthetic speech. We measured Root Mean Squared Error (RMSE) and Pearson's correlation, and present results in Table 5.1. The second objective evaluation compared the $F_0$ values directly, again using RMSE and Pearson's correlation, with results presented in Table 5.2. Additionally, the $F_0$ contours of an example utterance produced by each system are found in Figure A.3 in the appendix.

As hypothesised, **ORACLE-GT-F0** performs the best across all metrics. **LEGENDRE** outperforms both **BASELINE** and **CATEGORICAL** in all metrics, achieving higher correlations and lower RMSE for both polynomial coefficient values, and for $F_0$ values. This indicates that our hierarchical Legendre polynomial model offers more accurate control (in the current work, this was an intonation transfer task) than conventional pitch accent label conditioning.

The much higher correlations of polynomial coefficient and slope values for **LEGENDRE** than **BASELINE** suggest that **LEGENDRE** has indeed learned to use the provided conditioning coefficient values to predict a more accurate $F_0$ contour. Some of the difference in $F_0$ RMSE (Table 5.2) between **ORACLE-GT-F0** and the **LEGENDRE** model might be because **LEGENDRE** only receives non-zero coefficient values on prominent words, while **ORACLE-GT-F0** receives $F_0$ conditioning for every phone in every word.

Table 5.1: *RMSE (lower is better) and Pearson's correlation (higher is better) of each polynomial coefficients and slope values between each model and ground truth, over 50 test utterances.*

| **RMSE** (no units) ↓ | | | | |
|---|---|---|---|---|
| **Model** | **Leg-0** | **Leg-1** | **Leg-2** | **Slope** |
| **ORACLE-GT-F0** | 0.72 | 1.25 | 1.50 | 0.76 |
| **BASELINE** | 1.90 | 2.79 | 2.62 | 2.40 |
| **LEGENDRE** | 1.00 | 1.58 | 1.60 | 1.31 |
| **CATEGORICAL** | 1.94 | 2.79 | 2.89 | 2.35 |
| **Pearson's correlation** ↑ | | | | |
| **Model** | **Leg-0** | **Leg-1** | **Leg-2** | **Slope** |
| **ORACLE-GT-F0** | 0.927 | 0.914 | 0.845 | 0.950 |
| **BASELINE** | 0.450 | 0.471 | 0.428 | 0.470 |
| **LEGENDRE** | 0.860 | 0.859 | 0.808 | 0.870 |
| **CATEGORICAL** | 0.477 | 0.473 | 0.327 | 0.488 |

Table 5.2: *RMSE and Pearson's correlation of $F_0$ between each model and ground truth, over 50 test utterances.*

| **Condition** | **RMSE (**Hz**)** ↓ | **Correlation** ↑ |
|---|---|---|
| **ORACLE-GT-F0** | 63.19 | 0.838 |
| **BASELINE** | 73.47 | 0.775 |
| **LEGENDRE** | 68.65 | 0.808 |
| **CATEGORICAL** | 72.35 | 0.784 |

## 5.5 Discussion

The goal of this chapter was to provide intonation control in a speech synthesis model. Such control has two main benefits. First, it can allow control via external prediction models which can be trained on data which may not be suitable for training speech synthesis models, for example conversational data. Second, controllable synthesis can be

used to create stimuli for perception experiments in prosodic research (see Chapter 6). In particular, this can benefit research into conversational speech, as spontaneous conversational speech is difficult to elicit in a controlled manner using lab speech. Furthermore, the method in this chapter could be used to perform voice puppetry, similar to the work in Van De Vreken et al. (2022), by extracting the Legendre polynomial features from a reference voice, and using these features to control the synthesis of another voice.

In the literature review in this chapter, we saw that there are many ways to add prosodic controllability to a speech synthesis model, using representations which can acquired in multiple ways. One of the main goals of this chapter was to use an interpretable and linguistically-relevant representation. To achieve this goal, we modelled intonation using parametric representation of $F_0$, and unlike many forms of prosodic control, for example Global Style Tokens (Skerry-Ryan et al., 2018), we wanted to model this on a more linguistically-relevant level, namely the word- and phrase-level. We took inspiration from the CoPaSul model, which had not yet (to our knowledge) been implemented in a speech synthesis model. This meant using a superpositional approach, modelling phrase-level features using a linear slope and word-level prominence shape using Legendre polynomial coefficients. We chose Legendre polynomials due to their interpretability and orthogonality. Further, they are well-motivated having been used in previous prosodic research to validate human-annotated data based on the autosegmental-metrical framework. We compared this approach to a categorically-conditioned model, where words received a categorical prominence or boundary label.

In the subjective evaluation, which probed perceptual similarity, we found that the **LEGENDRE** model significantly outperformed the **BASELINE** model, which did not receive any external conditioning during inference, and the **CATEGORICAL** model, which only received information about the presence of prosodic prominence and boundaries. This suggests that our model provides more information about $F_0$ than simply information about the presence of prominent words and boundaries. This was verified using an objective evaluation, which measured the RMSE and correlation between the Legendre coefficients and slope of the reference utterances and the resulting speech synthesis output. We also measured the RMSE and correlation across the entire $F_0$ contour, and these metrics showed similar results. This was expected, because simply providing boundary information or prominence information during training will not provide information about whether a prosodic boundary is rising or falling, for example. Because the categorical features can captures more information than $F_0$, for example duration and energy, a future model could use both categorical features in conjunction with the sparse representation of $F_0$ from the **LEGENDRE** model.

Our approach, however, did not beat the **ORACLE-GT-F0**. We hypothesised that this would be our topline. This is due to the fact that this model receives $F_0$ values for each individual phone making it more granular and precise. This model outperformed the **LEGENDRE** model in both the subjective evaluation and all of the objective measures. The **LEGENDRE** model did perform more closely to the **ORACLE-GT-F0** model than the other models, with a high correlation and lower RMSE on all objective measures. This suggests that the **LEGENDRE** model does learn the information about the coefficients and slope and that it does provide adequate control. Further, the **LEGENDRE** model did have a certain disadvantage compared to the **ORACLE-GT-F0** model in that only words which were found to be prominent received coefficients. Due to how we discretised the prominence values, this means that 47% of words received no conditioning. A potential

improvement would be to condition all words using polynomial coefficients. Furthermore, the **ORACLE-GT-F0** error could be improved by conditioning the model on $F_0$ per mel spectrogram frame, as is done FastSpeech 2 (Ren et al., 2021), rather than per phone. Modelling $F_0$ per input symbol removes information about the $F_0$ trajectory on individual segments, such as the vowels. Therefore, future work should compare the methods proposed in this chapter, to a model, such as FastSpeech 2, which can provide an even more granular representation of $F_0$ to condition the model.

The results suggest there is a trade-off between the granularity of a representation and the interpretability of a representation. The **LEGENDRE** model potentially allows for more interpretable and linguistically-relevant control due to its sparse representation of the same information. For example, to change the slope on a particular pitch accent in the **ORACLE-GT-F0** model it would require a user to know per phone which $F_0$ value to change in order to achieve that slope globally on the word. In our model, this can be achieved by changing the value of the second Legendre coefficient. Using a word-level representation is therefore more user-friendly and mirrors how word-level information is described in prosodic research. However, to investigate the proposed model's usability as a human-in-the-loop model, more research is needed.

There is an additional benefit of using a word-level, rather than phone-level feature for $F_0$. Many external models that predict prominence, operate on the word-level rather than the phone-level, and our model provides a single word-level representation to enable this. This could allow us to create predictive models using features like word embeddings, such as BERT embeddings, mirroring the studies of Talman et al. (2019); Kakouros and O'Mahony (2023). This would also allow enable training predictive models using conversational data that is not suitable for training speech synthesis models (Ben-David and Shechtman, 2021).

Further, our representation has been found to be related to more traditional prosodic annotations, such as ToBI labels, and therefore is a more linguistically-motivated representation than a phone-based representation. This can allow us to characterise pitch accents using their shape, which is common in linguistics. In Chapter 6, we will present a case study in which we use these representations to explore the realisation of a discourse marker in a corpus of spontaneous speech and subsequently use the findings from this exploration to create stimuli using the model presented in the current chapter.

There are however some disadvantages to the method presented in the current chapter compared to creating stimuli using post-processing techniques such as TD-PSOLA. When a model is *conditioned* on features, such as coefficients or indeed $F_0$ values per phone, we may not achieve the exact intended $F_0$ realisation in the resulting speech output. This is expected, as conditioning does not enforce the intended $F_0$ information onto the speech, but rather conditions the model to perform in a certain way. This is why we do not see a perfect correlation or RMSE for $F_0$; while when using post-processing techniques, this can be achieved. At the same time, when using conditioning, the model should behave in a way that reflects the training data. If we were to enforce a very unlikely contour onto an audio sample using TD-PSOLA, we might achieve that contour, but with many artifacts, making the speech sound unnatural or distorted. When using conditioning, we might expect the model output to reflect the constraints imposed by the training data. This can be seen as a positive and a negative, but the result is that the speech from a conditioned model should sound less distorted than it would using post-processing. As mentioned, future work should investigate conditioning the model on frame-level $F_0$ to evaluate whether this improves the accuracy of $F_0$ synthesis.

Additionally, for validating the use of this method in speech science research, it would be helpful to compare the results of this method to human performance in prosodic imitation experiments. For example, Chodroff and Cole (2019) investigated the prosodic realisation of eight nuclear intonation tunes in English across speakers. To do this, they resynthesised short utterances by imposing one of eight intonational contours on the final noun phrase. During the production experiment, participants listened to three utterances spoken with the same intonational tune after which the target sentence was presented. Participants were instructed to produce the target sentence with the same intonation and "said the way you think it should sound if it were spoken by a human English speaker, in a manner that is familiar to you." (Chodroff and Cole, 2019, p.1967). They found significant differences between the intended tune and tunes realised by their participants as calculated using RMSE between the intended contour and the realised contour with not all eight tunes being differentiated by speakers (Chodroff and Cole, 2019). To compare our method to human performance, the tunes in the experiment of Chodroff and Cole (2019) could be synthesised by extracting the slope and Legendre polynomial features and passing them to our model to compare to a human baseline for $F_0$ imitation.

Finally, we took a superpositional approach by using both a phrase-level slope coefficient and a word-level set of Legendre polynomial coefficients to condition our model. We also saw in Corkey et al. (2023) that good results can be achieved by using a single set of coefficients to condition phrase-level intonation. This suggests that this method can operate across multiple domains. This means that this approach can be applied in many configurations, for example we could take a sequential approach by removing the additional phrase-level slope coefficient. Such an approach would be more similar to the PaIntE model. We could also use all four coefficients from the third-order Legendre polynomial, recall here we took the first three coefficients. Adding the final coefficient would add more detail to the word-level representation, though we did see in the objective evaluation that the correlation for Leg-2 is lower than the first two coefficients (Leg-0 and Leg-1), which suggests that learning more detailed aspects of the $F_0$ contour might be more difficult. Adding more than four coefficients is also possible, but from that point, the coefficients are not easily interpretable and though they will provide more detail, they may only be learning micro-prosodic information, which may not generalise across words.

## 5.6 Conclusion

In this chapter, we have shown that using a data-driven hierarchical specification of the $F_0$ contour on the phrase-level using slope and on prominent words using Legendre polynomial coefficients outperforms traditional binary categorical conditioning: our **LEGENDRE** model produces an $F_0$ contour that sounds more similar to the reference $F_0$. As expected, our **LEGENDRE** model did not beat the **ORACLE-GT-F0** model, but our model was at a disadvantage because **ORACLE-GT-F0** received $F_0$ values for both prominent *and* non-prominent words. In future work, we could provide polynomial coefficients for all words, obviating the need to first detect prominent words.

Our proposed method would also allow for $F_0$ transfer from a *different* speaker, potentially allowing us to transfer conversational intonation patterns from low-quality recordings onto a speech synthesis model speaker for which we have no conversational speech. In Chapter 6, we will use the representations from the current chapter to explore

patterns in a corpus of conversational speech. We will subsequently use the patterns found in the exploration in Chapter 6 to control the speech synthesis model presented here to create stimuli for a perception experiment investigating the effect of prosody on the level of agreement or disagreement expressed by a speaker.

# Part IV

# Case Studies

# 6

# Exploring the Prosody of Discourse Markers Using Found Data and Speech Synthesis

This chapter is based on the following paper:

**O'Mahony, J.**, Lai, C., & Székely, É. (2024). "Well", what can you do with messy data? Exploring the prosody and pragmatic function of the discourse marker "well" with found data and speech synthesis, *Proc. Interspeech 2024* (pp 4084 - 4088), Kos, Greece

The first author was responsible for creating the corpus of discourse markers from found data, clustering and analysing the data, training the TTS models, creating the stimuli for the listening experiments, running the evaluation and analysing the results. The contributions of the second author were supervisory in nature, and include verifying statistics and editing the paper. The contributions of the third author was also supervisory in nature, including editing and providing the tSNE figure. The chapter below contains significant textual overlap with the article above.

## 6.1 Introduction

In the previous chapters, we introduced speech synthesis methods that can allow us to study prosody in conversational speech. In Chapter 4, we addressed the data sparsity problem by using found data to improve the realisation of a group of speech acts that are often underrepresented in read speech, namely questions. This method illustrated the potential of using found data to improve the realisation of certain conversational structures which are not abundantly present in read speech datasets. But synthesising questions by adding exemplars to the training data will not allow us to synthesise the immense

*variability* in their realisation, variability which is caused by various contextual features, such as a speaker's attitude or stance, discourse factors, such as turn-taking, or socio-indexical features (Ogden, 2006, p.1753).

In Chapter 5, we therefore introduced a method for controlling intonation with a sparse word- and phrase-level representation of $F_0$ which can be extracted from found data, and in particular data which may be sub-optimal for use in speech synthesis modelling. By using a sparse representation that can control a TTS model, we can also use this representation to characterise prosodic patterns in noisy found data and use these patterns to control a speech synthesis system to study their perceptual relevance. In this chapter and the next chapter, we present two case studies on how we can use speech synthesis methods to study conversational speech. In particular, we focus on communicative features at the beginning of utterances, and the end of utterances. In Chapter 7, we will explore how speech synthesis models can be used to synthesise and analyse turn-taking cues found utterance-finally. But first, in the current chapter, we use the methods previously described to explore the prosodic variability of the utterance-initial Discourse Marker (DM) *well*, and one of its pragmatic functions in a corpus of found data.

As we discussed in Chapter 2, one of the benefits of using found data is that we can use greater quantities of data to capture a broader range of prosodic variation, as well as conversational speech phenomena such as backchannels (Figueroa et al., 2022), filled pauses (Székely et al., 2019a; Dall et al., 2014a), discourse markers (Wester et al., 2015; Andersson et al., 2012). Furthermore, using larger quantities of data allows us to capture a greater number communicative contexts in which utterances occur. However, the benefit of having an extreme amount of variation to model, simultaneously presents a huge challenge, namely *accounting* for that variation in the data (Ogden, 2007).

At the most fundamental level, we ideally need information about the speakers and their sociolinguistic backgrounds. This is because different sociolinguistic groups can show systematic differences in their prosody (Grabe et al., 2005). But found data often comes without metadata containing basic information on a speaker's sex, dialect, or linguistic background – information which would be available in a linguistically-curated corpus. Further, found data can exhibit a wide range of recording conditions and communicative contexts in which the speech occurs (e.g. a casual conversation versus a formal interview). Thus, next to the pragmatic aspects we may wish to model, we are also confronted with sociolinguistic and extralinguistic variation. Because all of these features share the same signal (Ogden, 2006), disentangling sources of variation can be extremely difficult, if not impossible. Accounting for this variation is desirable because it might also exert influence on the realisation of pragmatic functions.

Unfortunately, automatically accounting for the variation in speech data is difficult. We either need human annotations, which are time-consuming and expensive to generate, especially when using large quantities of found data (Lyth and King, 2024), or we can use models trained to identify particular characteristics of the data. An example of the latter approach was proposed by Lyth and King (2024), who used various models to label training data for aspects of audio quality such as SNR and C50 (a measure of reverberation), dialect, and global $F_0$ information. These are of course mainly extralinguistic and indexical features, but nonetheless remove confounding sources of variation in the data.

For aspects of variation which are affected by pragmatics and context, however, we do not have straightforward methods to label data automatically, and moreover there are few

datasets on which we can train classification models. This is because many aspects of pragmatic meaning in conversation are extremely nebulous, and as noted by (Ogden, 2006, p.1753): "terminology used to label such things as 'speaker attitude' can be highly subjective. What one person labels 'energetic' may be labelled as 'aggressive' by someone else: such terms have no clear empirical warrant other than the analyst's intuition as a native speaker". But without accounting for variation in the data and the pragmatic sources of this variation, we risk synthesising *average prosody* (Hodari et al., 2019). Similarly if we neglect to model prosodic outliers, or infrequent prosodic patterns, we will potentially ignore salient prosodic structures which may signal pragmatic functions (Ogden, 2012; Cruttenden, 1984; Freeman et al., 2015b). A step we can take towards exploring the prosody-pragmatics interface is to apply quantitative methods to conversational found data to *identify* commonly occurring patterns and to subsequently use speech synthesis models to synthesise this variation and evaluate its pragmatic effect.

In this work, we propose a method to understand the variation in an unlabelled dataset of conversational speech and in particular how this variation might be related to pragmatic meaning. We use unlabelled data, in contrast to an annotated corpus such as Switchboard NXT (Calhoun et al., 2010), to explore whether such unlabelled data can provide an additional source of information for phonetic studies. Here, we focus on DMs, which have received relatively little attention in speech synthesis research, but which make up a significant subset of the top twenty most frequent words in conversational English (Andersson et al., 2012). Their frequency in conversation is due to their multi-functionality, especially in signalling the relationship between prior and present turns in conversation (Yang, 2002). For the purposes of this study, DMs are an interesting case of the pragmatics-prosody interface as their function can be related to their prosodic realisation (Hirschberg and Litman, 1993; Lee et al., 2020; Yang, 2002). From a practical perspective, DMs are easily identifiable in corpora due to their fixed lexical content[1], and due to their importance in contextualising turns in a conversation, they are a useful point to begin explorations of contextualised speech (Yang, 2002). In this study, we explore the prosodic variation of the DM *well*, one of the most studied discourse markers in English (Heritage, 2015). *Well*, also consists solely of voiced segments and a single syllable making it an ideal candidate for extracting prosodic features, such as $F_0$.

To investigate how we can use unlabelled found data to study the prosody-pragmatics interface, we present a small case study on the relationship between the prosodic realisation of the discourse marker *well*, and the pragmatic effect of the prosodic realisation. Specifically, we study how the prosodic realisation of the DM can affect the level of *agreement* expressed by a speaker. In this chapter, we therefore ask the following research question:

**RQ6.1** Does the level of agreement expressed by the speaker change as a function of the prosody of the Discourse Marker *well*?

To investigate the impact of prosody on the level of agreement/disagreement expressed by a speaker, we carry out the following methodology:

1. We develop a method of extracting a representative corpus of discourse markers from unlabelled speech data, for the example of *well*.

[1]We note here, however, that some DMs, such as *well*, have both a DM and non-DM meaning. The lexical form of DMs nonetheless allows us to easily search in corpora and narrow down the search space.

2. We explore the prosodic realisation of *well* in a data-driven manner by clustering the intonational representations introduced.

3. We synthesise prosodic variation on the discourse marker *well* based on the cluster centroids using controllable TTS.

4. We evaluate the effect of the prosodic realisation of *well* on the level of *agreement/disagreement* perceived by listeners.

The overarching goal of this study is to illustrate the above method of studying conversational speech for a very specific case study. However, we present this method as a general approach which can be applied to study the prosody of any discourse marker in conversation. This work has four potential applications relating to speech synthesis and linguistic and phonetic research.
Speech synthesis research:

1. From the perspective of training speech synthesis models, this method can allow researchers to identify structures in the found data that they may wish to model. Due to the large number of rare events (LNRE) in speech corpora (Möbius, 2003), it can allow us to detect infrequent, but potentially pragmatically salient patterns in found data.

2. For speech synthesis evaluation, this method can allow researchers to select a wide range of realisations to be included in listening experiments and to potentially identify specific contexts in which these realisations occur.

Linguistic and phonetic research:

3. For pragmatics research, which often concentrates on qualitative analysis of a limited number of conversational excerpts, this method can allow for the identification of structures which may warrant further investigation using larger quantities of data.

4. For experimental prosodic research, this method can allow us to use acoustic patterns found in real data to create experiments (as we will see in this chapter), rather than manipulating stimuli with top-down knowledge.

## 6.2 Related Work

### 6.2.1 Discourse Markers

Discourse Markers are a class of words, such as *oh*, *so*, and *well*, or multi-word phrases, such as *you know* or *I mean*, which are ubiquitous in conversational speech[2] (Aijmer and Simon-Vandenbergen, 2003; Popescu-Belis and Zufferey, 2011) and make up many of the lexical entries in the top twenty most frequent words in conversational corpora (Andersson et al., 2012). They are described as "semantically empty" (Brinton, 1996, p.35) expressions, that is to say, they do not alter the propositional meaning of an utterance, but as Brinton

[2]They are also used in written language and spoken monologues, but for the purposes of this chapter we will focus on their use in conversational speech.

continues "they are not pragmatically optional or superfluous: they serve a variety of pragmatic functions" (Brinton, 1996, p.35). They are particularly involved in structuring discourse (Hirschberg and Litman, 1993) and contextualising or conveying the relationship between utterances (Yang, 2002; Blakemore, 2002; Brinton, 1996). Because of their importance in conversation, they have been well-studied in the fields of semantics, pragmatics and CA, from both a quantitative and qualitative perspective. Further, because of their importance in signalling discourse structure they have also received interest from an Natural Language Processing (NLP) perspective, especially in relation to discourse parsing and segmentation (Hirschberg and Litman, 1993; Popescu-Belis and Zufferey, 2011).

### 6.2.2 Previous work on *well*

In this chapter, we present a case study of one of the most frequently-studied DMs, *well* (Heritage, 2015). The use of *well* carries many functions (for example, Brinton (1996) lists 15 in total), including but certainly not limited to connecting past to present utterances (Bolden, 2006; Blakemore, 2002), shifting topics (Aijmer, 2015), holding the floor (Brinton, 1996), conveying alignment between speaking partners (Bolden, 2006), expressing stance (Sakita, 2017), signalling that an answer is insufficient or dispreferred (Sakita, 2017; Heritage, 2015), hedging a disagreement (Helt, 1997; Michilsen, 2019; Innes, 2010; Aijmer, 2015), or signalling agreement (Brinton, 1996; Innes, 2010). Due to the multi-functionality of this DM, and its role in signalling sometimes opposite functions, for example agreement but also disagreement, it is no surprise that its prosodic realisation has been studied and has been found to differ depending on its function.

Hirschberg and Litman (1993) investigated the role of prosody in the disambiguation of the function of *well*. Here they were not interested in the aforementioned pragmatic functions of *well*, but in correctly identifying when it was used as a discourse marker as opposed to its sentential meaning as an adverb. They used symbolic representations of pitch accents and phrasing in their prosodic analysis. They studied 52 examples of *well* of which 27 were found to be used as a discourse marker. They found that when used as a discourse marker, just over half of the examples formed single prosodic phrases, and the remainder of the samples were found in the first position of a prosodic phrase with the majority of phrase-initial realisations receiving no pitch accent. In their study however, no acoustic measurements were used to examine finer phonetic details of the realisations. Further they used very few examples, making the generalisability of the results weak.

Popescu-Belis and Zufferey (2011), similarly explored the differences between the sentential and discourse marker use of *well* using decision trees to classify each function. They used a larger number of samples than the study of Hirschberg and Litman (1993), 4,136 instances, all of which were annotated. They found that 88% of the instances of *well* functioned as a discourse marker. They used sociolinguistic, textual, prosodic, and positional features in their classifier. Unlike Hirschberg and Litman (1993), they did not use any intonational information in their set of prosodic features, but instead looked at duration of the tokens and surrounding pauses. They found that the only prosodic feature which improved classification performance was the presence of a pause before and after the discourse marker. They acknowledged that this feature potentially encoded positional information, which was also found to be highly relevant due to the fact that DM function of *well* most often occurred in utterance initial position in their study: in roughly 75.5% of cases. The presence of a pause following the DM would also potentially indicate that the

DM is part of its own prosodic phrase as was found in the study of Hirschberg and Litman (1993) and has been noted in later studies such as Lam (2009). Both of the studies above, however, neglected to study the specific pragmatic functions of *well*. The lack of significance of duration in Romero-Trillo (2018) in disambiguating the sentential from discourse marker form may reflect the fact that this DM performs a large number of functions, which themselves might show differences in duration.

In contrast to the studies above, both Romero-Trillo (2018) and Michilsen (2019) carried out investigations into the prosodic realisation as a function of various pragmatic functions of *well*. Romero-Trillo (2018) used a symbolic representation of pitch accents to characterise the intonation of *well*, specifically characterising the rises and falls of the pitch accent. Further, he included a characterisation of position of each DM within a prosodic phrase. He found that the majority of instances of *well* in his study, were used as discourse markers (442 with a DM function vs. 60 non-DM function), mirroring the results of Popescu-Belis and Zufferey (2011). The majority of instances of *well* were realised with no prominence (339), followed in frequency by a prominence with a falling tone (70). He found that the disagreement function of *well* was most often realised with no pitch accent, followed in frequency of occurrence by pitch accents with a falling tone, however he found evidence for *well* being realised with all pitch accent types in his corpus. A total of 37% were used for the function of disagreement. Further, mirroring the results of Hirschberg and Litman (1993), he found that the word was found 271 times in phrase-initial position, and 73 times comprising of a unique prosodic phrase. Romero-Trillo (2018) postulated that the importance of the initial position of *well* is important for contextualising the following utterance.

Michilsen (2019) used a smaller dataset comprising of 174 instances of *well*. She investigated the relationship between the pitch accent type, the level of phonetic reduction reduction, duration, and phrase and turn position as a function of four functions of *well*, namely *marker of insufficiency*, *face-threat mitigator*, *frame*, and *delay device* (Michilsen, 2019, p.9). As in the study of Romero-Trillo (2018), she used a symbolic representation of intonation following the autosegmental metrical theory. She found that 57.3% of the instances of *well* occurred turn-initially, and 69.5% of the total instances received no pitch accent. For the function of face-threat mitigator, which signals disagreement, the mean duration was significantly higher than for other functions. For pitch accents, she found no significant difference between the different functions of *well* and like Romero-Trillo (2018) found that the majority were unaccented. While the studies of Romero-Trillo (2018) and Michilsen (2019) suggest that *well* is likely to be deaccented, this is not always found to be the case. In a study by Lam (2009) of *well* in Hong Kong English, the DM was found to occur in the majority of cases as an isolated prosodic phrase which was always accented, while the deaccented form was only found in 31.6% of instances. Again, like the study of Romero-Trillo (2018), the phonetic realisation of $F_0$ was not included in the study and in both studies relatively few examples were used.

In the studies mentioned above we see different approaches to characterising the prosody of *well*. None of the studies used quantitative measurements of $F_0$ or appeared to take the position of the discourse marker in the speakers' $F_0$ range into account. Where a symbolic characterisation of pitch movement was used, there appears to be a strong tendency for *well* to be realised with no accent, however two of these studies used under 200 samples, which may affect generalisation. Across all studies, we see that *well* is most often used as a discourse marker and when it is used as a discourse marker it is found

phrase-initially or turn-initially. Further, the presence of silences before and after the token suggests that it may occur utterance-initially when using IPUs (as we will do in this study), and that is can often occur within its own prosodic phrase. Finally, none of the studies above used perceptual experimentation, thus the importance of prosody for the perception of this DM has not been established.

### 6.2.3 Clustering Prosodic Features

As mentioned in Section 6.1, we propose to cluster prosodic patterns in a corpus of found data to identify common patterns and test their perceptual relevance with regard to the degree of agreement/disagreement perceived by listeners. We are not the first to propose clustering prosodic features to explore pragmatic meaning. Zellers and Ogden (2014), for example, used clustering in conjunction with a conversational analytical approach to examining the prosodic features of target utterances conveying contrast in relation to their preceding utterance, the context utterance. They found pragmatic differences between the context-target pairs which were realised with similar prosodic features (i.e. were found in the same cluster), to context-target pairs which differed in their prosodic features. In this study, however, they only used 27 examples because they were limited to data which had been qualitatively annotated in a conversational analytical framework. For the purposes of developing TTS systems and using large amounts of found data, the need for detailed transcriptions will not scale. Furthermore, they did not perform perceptual experiments to evaluate the role of prosody on the perception of contrast.

Calhoun and Schweitzer (2012) similarly used clustering to investigate evidence of lexicalised prosody. To do this, they parametrised the $F_0$ contour of words and collocations containing an annotated nuclear accent in the Switchboard NXT corpus (NXT) (Calhoun et al., 2010) using the PaIntE parameters, which we briefly introduced in Section 5.2.2. Here they investigated whether frequent collocations have lexically-stored intonation, which would be evidenced by frequent collocations sharing similar intonational features. To investigate this they clustered PaIntE parameters and compared the clusters to hand-labelled pitch accents from the NXT corpus. They found that collocations within the same clusters often contained a specific discourse meaning, and that these collocations most often contained a discourse marker. Interestingly, they found that there was little overlap between the clusters of tokens which received the same symbolic nuclear accent in the transcription, which further emphasises the point that a phonetic representation of the $F_0$ contour can contain more potentially discriminative information than a symbolic representation (Calhoun and Schweitzer, 2012).

Though they were not specifically interested in particular discourse functions, they did conduct a perceptual experiment to test whether utterances containing phrases that were found to have a high-frequency of similar intonational characteristics were deemed to be more acceptable than low-frequency equivalents. They found that participants deemed the high-frequency condition to be more acceptable in a rating experiment. To create the stimuli they used extracts from Switchboard NXT, however, to keep the speaker constant, two speakers were recorded reenacting the extracts. The $F_0$ of these reenactments were manipulated in Praat and the authors noted that in 33% of the stimuli, post-processing the $F_0$ in the recordings lead to distorted speech. This again shows that for intonational research, synthesising patterns with TTS may be a better option, which would alleviate the need to rerecord stimuli, and offer control over the prosodic aspects of the speech, without

too much signal distortion. The method presented in this chapter, similar to that of Calhoun and Schweitzer (2012), will use a parameterised representation of intonation, which should contain more detail than the symbolic descriptions presented in the previous section. Unlike the study of Calhoun and Schweitzer (2012), however, we will use controllable TTS to generate our stimuli.

## 6.3 Method

### 6.3.1 Curating a Discourse Marker Corpus

#### Conversational Data

We used the CANDOR Corpus (Reece et al., 2023), which contains 1656 dual-channel recordings of conversations lasting roughly 25 minutes. The conversations were recorded via internet on a video-call platform and take place between two strangers. Though this corpus was curated by Reece et al. (2023) for scientific purposes, we characterise this as found data because the corpus was automatically transcribed using Amazon Web Services (AWS) ASR and therefore contains no ground-truth transcription or additional linguistic annotation. Various speaker demographic information was collected such as age, educational attainment, and sex. The corpus contains a wide range of recording conditions (probably due to varying quality microphone and internet connection), background noise, and occasional channel leakage if speakers were not using headphones.

#### Data Filtering

Unlike the The Spotify 100 000 Podcast Dataset used in Chapter 4, the CANDOR Corpus is dual-channel, so to extract utterances, we split each channel into interpausal-units (IPUs) delimited by a silence threshold of 180 ms. Frequent one word backchannels were removed, and the IPUs were classified following the communicative state classification in Heldner and Edlund (2010). We extracted all IPU starting with the word *well*. We chose to only consider IPU-initial tokens based on the research described in Section 6.2.2, which found the position within an utterance and prosodic phrase to be highly predictive of the token functioning as a DM. Because a speaking turn can comprise of multiple IPUs, we classified each IPU depending on its position within a turn and kept IPUs which were turn-initial, turn-medial, or an overlap initiated by the speaker.

Due to the varying audio and transcription quality in the dataset, we ran a number of filtering steps. We first used a better performing ASR model, Parakeet[3], to re-transcribe each IPU in the dataset. To verify the presence of utterance-initial *well*, we then compared the new transcription with the original ASR transcript and kept IPUs if both transcripts agreed on the presence of the initial *well* token. To estimate SNR of each recording, we used Brouhaha (Lavechin et al., 2023)[4] and removed IPUs with an SNR of 25 or lower to filter some of the most noisy data.

Each IPU was then aligned using the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) and IPUs where forced alignment failed were removed. We observed that many of the word alignments for our word of interest also contained the second word in the utterance.

[3] `https://huggingface.co/nvidia/parakeet-tdt-1.1b`

[4] `https://github.com/marianne-m/brouhaha-vad`

This is due to the nature of the data and the fact that alignments for spontaneous speech may contain more errors than read speech (Dall et al., 2016a). We therefore spliced out the word of interest *well* and ran the spliced audio through Parakeet ASR in a second pass to verify that the spliced word was correct and didn't contain additional words[5]. Finally, we used the presence of silence after the *well* token as a proxy for a prosodic boundary, as in this study we are interested in tokens which occur in their own prosodic phrase. Details of IPUs remaining after each filtering step and final data quantities can be found in Table 6.1. While a significant amount of the potential data is lost, a trade-off has to be considered between quantity and quality when using found data and we chose a conservative approach. In total, there are 484 speakers in the the *well* dataset.

Table 6.1: *Remaining IPU quantity after each filtering step*

| Filter | Number of IPUs |
|---|---|
| **Raw** | 8095 |
| **Parakeet Errors** | 8067 |
| **Parakeet match to ASR** | 5780 |
| **SNR** | 5319 |
| **MFA Errors** | 5262 |
| **Parakeet on spliced *well*** | 1883 |
| **Post-*well* silence** | 942 |

### Acoustic Information

To explore the prosodic realisation of the discourse marker *well*, we extracted a number of features relating to the $F_0$ contour, as well the duration of each realisation. Global $F_0$ information was first extracted to normalise the word-level $F_0$ features. To do this, we the extracted the $F_0$ per speaker across every conversation that they took part in, using Parselmouth (Jadoul et al., 2018) with the Praat autocorrelation $F_0$ extraction function (Boersma and Weenink, 2021).

After splicing the *well*-tokens from each IPU, word-level $F_0$ was extracted using the Praat filtered autocorrelation function in a two-step manner following de Looze and Rauzy (2009). We used the Praat "kill octave jumps" function and interpolated the $F_0$ values. $F_0$ was converted to semitones relative to the *global* speaker median. $F_0$ values were then $z$-score normalised using the *global* speaker mean and standard deviation in semitones. We chose to normalise by global features to capture where the *well* realisation fell in a speaker's global range, which may be pragmatically meaningful, and normalising the $F_0$ on the utterance-level would not capture this.

To describe the $F_0$ characteristics of each instance of *well*, we used a third-order Legendre polynomial to model each $F_0$ contour. This representation was used in the previous chapter because of its orthogonal and interpretable coefficients. Recall that the first three coefficients characterise the $F_0$ height (LC0), slope (LC1), and curvature (LC2)

[5]We acknowledge that this can lead to potential issues because ASR will often perform worse when transcribing single words, but since we are solely interested in the realisation of single words, this step was necessary. Future improvements in alignment models for spontaneous speech would increase data retention.

respectively. To extract the polynomials, we time-normalised the $F_0$ contour between -1 and 1, and fit a third-order Legendre polynomial and extracted the first three coefficients. Finally, we extracted the duration of each *well*-token according to the MFA alignments. We removed tokens if their duration fell below the first percentile or above the 99th percentile (15 removed) and one token failed during $F_0$ tracking.

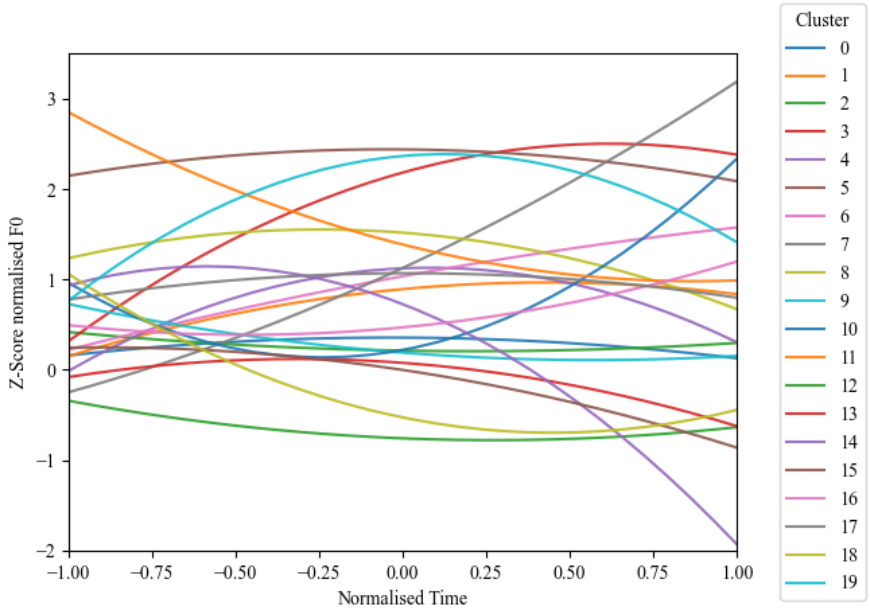


Figure 6.1: *Curves found through clustering and reconstruction of Legendre Polynomials*

### Clustering Prosodic Features

In this study, we were interested in the variation in the realisation of *well* in our dataset in terms of the $F_0$ contour and duration. To explore this variation, we used k-means clustering. As we saw, clustering has been used in previous research to find prosodic patterns in data (Zellers and Ogden, 2014; Reichel, 2007; Calhoun and Schweitzer, 2012). By partitioning the acoustic space into clusters of similar prosodic features, we could then use the cluster centroid, or average of each cluster, to control a TTS model. For our clustering we used four features: the duration of each token in seconds, $F_0$ height (LC0), slope (LC1), and curvature (LC2). This led to a feature vector of four dimensions for each of our 926 *well* samples.

To cluster the data, we used the k-means algorithm from the Python package sklearn (Pedregosa et al., 2011). We scaled each dimension using sklearn RobustScaler which scales the data based on percentiles and is robust to outliers. This allowed us to retain values at the tails of our distribution in the clustering, because prosodic outliers may be perceptually meaningful. We chose to split the feature space into 20 clusters to cover enough variation in the acoustic space for stimuli creation for our experiment and because there is no theoretical optimal number of clusters for this task. This study is therefore explorative in nature. We do not, however, assume stringent categories of prosodic realisations, but use this as a tool to partition groups of closely-related features.

Results of the clustering are visualised using tSNE in Figure 6.2 and the reconstructed polynomials of each cluster centroid can be seen in Figure 6.1. The tSNE visualization

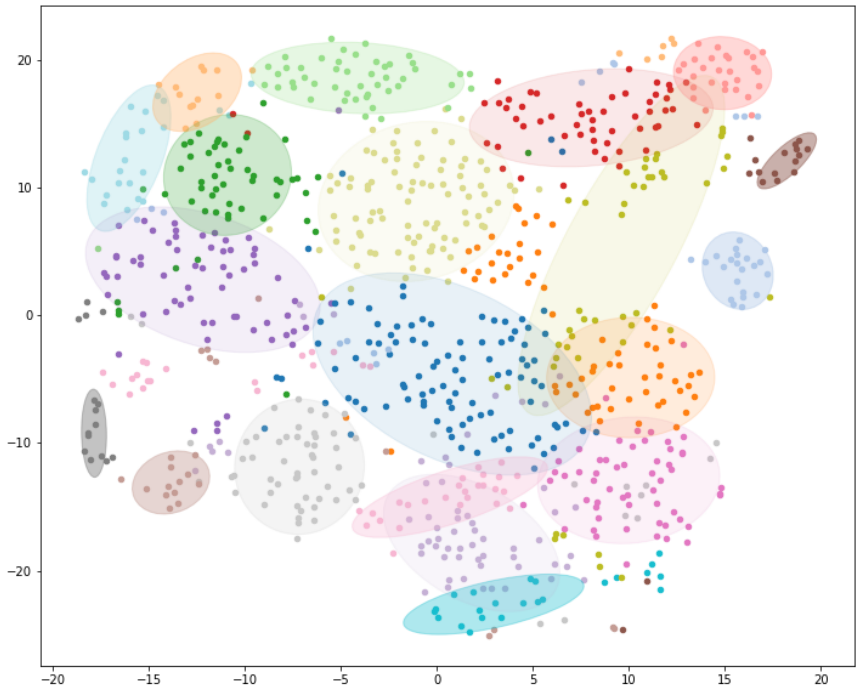


Figure 6.2: *tSNE plot showing clusters found in k-means (made by co-author Éva Székely)*

reveals that while some clusters, especially those at the center of the feature space, overlap, peripheral clusters are smaller and more distinct. Such overlap is expected, given the dataset's diversity, and this is common when clustering prosodic features (Calhoun and Schweitzer, 2012). Nonetheless, informal auditory evaluation of samples closest to the cluster centroids revealed a perceptual similarity within each cluster.[6].

### 6.3.2 TTS

#### TTS Model

We used FastPitch (Łańcucki, 2021), which was described in each of the previous chapters. We provide additional conditioning parameters to our model, as described in Chapter 4. Recall that phrase-level slope and word-level Legendre polynomial coefficients can be used to condition the model hierarchically to specify the shape of the $F_0$ contour. The slope parameter is found by performing linear regression on the $F_0$ contour of each phrase in an utterance. This slope value is then upsampled to the number of phones in that phrase. For word-level Legendre polynomial features, the $F_0$ is de-trended based on the slope and the resulting $F_0$ curve of each prominent word in an utterance is found by fitting a third-order Legendre polynomial and taking the first three coefficients. Therefore at synthesis time, the user can specify a word-level accent shape by changing the coefficients of the input data. Specifically, for our purposes, this method enables the utilisation of the Legendre coefficients from the centroids of each cluster, along with their durations, to generate specific prosodic renditions of *well*.

[6]We recommend listening to some of the cluster examples which can be found here: `https://johannahom.github.io/IS-2024/`

**Corpora and training**

In Chapter 4, we found a benefit of mixing conversational speech with read speech: using read speech helps to create a stable synthetic voice, while the addition of spontaneous data exposes the model to conversational prosodic patterns. We therefore trained a model for this study using both conversational data and read speech data. For our read speech, we used the LJ Speech corpus (Ito and Johnson, 2017) consisting of 13 100 utterances of audiobook data read by a female US English speaker.

For our spontaneous speech, we use the AptSpeech corpus, a publicly available multimodal, multi-party corpus of spontaneous conversational speech (Kontogiorgos et al., 2018). It consists of 15 interactions between a moderator and two varying participants playing a collaborative game. The recordings of the moderator have been annotated and segmented into breath groups following Szekely et al. (2019) and consist of a single male speaker. In addition to the moderator's spontaneous speech, the moderator recorded prompts in read speech style, which we also used during model training. A summary of the exact data used in training is in Table 6.2.

Table 6.2: *Data used for TTS model building*

| Dataset | Train | Val | Test |
|---|---|---|---|
| **Male Speaker (Read)** | 2165 | 4 | 27 |
| **Male Speaker (Conversational)** | 5194 | 21 | 49 |
| **Female Speaker (Read)** | 12693 | 25 | 278 |
| **Total** | 20052 | 50 | 354 |

We trained a multi-speaker TTS model using the corpora described in Section 6.3.2. Each speakers' training utterances were conditioned on a speaker embedding, and no additional conditioning parameter was given to the model to distinguish speaking styles (similar to the study in Chapter 4). Our model was trained on a single GPU with a batch size of 16 for 500 epochs. We used the the pre-trained NeMo Hifi-GAN vocoder[7] to generate all audio samples.

## 6.4 Evaluation

### 6.4.1 Goals

As mentioned in Section 6.1, the use of *well* serves multiple functions in conversation including, but not limited to hedging, signalling dispreferred answers, and agreement or disagreement. In this study we focus on *well* and its use to signal the stance of agreement/disagreement. Here we investigate how much weight the prosody of the discourse marker has when the context it precedes remains constant. We present a perceptual study in which our target word *well* is followed by the words *yes* (positive polarity) and *no* (negative polarity). Here we ask: Does the level of agreement perceived by the listeners change as a function of the prosody of *well*? To test this, we use the centroids of the clusters found in Section 6.3.1 to generate 20 stimuli per polarity type.

[7] `https://github.com/NVIDIA/NeMo`

### 6.4.2 Stimuli Creation

We use the Switchboard NXT corpus (Calhoun et al., 2010) consisting of 1126 conversations labelled for Dialogue Act Dialogue Act (DA) information. We use this corpus as a source of conversational speech to extract carrier utterances for the experiments described below. Here we extracted utterance based on the DAs labels and did not partition the data into IPUs. All DAs were extracted which began with *well*. We limited the DA categories to *answers*, *no*, *hedge*, *agree*, *yes*, *reject*, *affirm*, *neg*, *ans-dispreferred* and *maybe*, categories which may show agreement/disagreement. We only chose utterances whose exact textual content appeared two or more times (144 utterances), to ensure we had enough candidate utterances of various texts. From these utterances, we removed items which were unintelligible, contained laughter, background noise and truncated words. This yielded 72 candidates. Finally, we selected stimuli in which the *well* was followed by a pause leaving a set of 18 utterances. For the experimental stimuli, we chose two carrier utterances from the set of 18 utterances which signal agreement or disagreement lexically: *Well no* (negative polarity) and *Well yes* (positive polarity). We additionally used *Well I don't know* as a filler item to represent an utterance lexically signalling neither agreement nor disagreement (neutral polarity). We used all 18 utterances[8] synthesised with the prosodic features extracted from the NXT speakers as additional filler items in the experiment.

As mentioned, various prosodic features, such as duration and $F_0$ can be controlled in FastPitch. To create stimuli, we extracted the prosodic features (LC0-LC2 and duration) of our NXT carrier utterances and used these features to synthesise each text. Mismatches in intensity and pronunciation between the utterances were not an issue as the only information taken from the carrier utterances was the text, the phones (canonical dictionary lookup) based on MFA, and information about the $F_0$ contour (LC0-LC2), as well as the duration of the words following the target word *well*. In this study, we used two target carrier utterances in which the DM was followed by a single word. We did not take longer utterances from different speakers, as using $F_0$ and phone durations might have led to salient mismatches in speaker identity between stimuli. To create prosodic variation, we used the prosodic features from the closest sample to the cluster centroid in Section 6.3.1. We adapted the duration of the silence between each *well* and its following context to be approximately 13 frames long. The experimental stimuli therefore consisted of 20 positive-polarity, 20 negative-polarity utterances which were generated using 20 *wells*, one from each cluster centroid. The fillers consisted of 20 neutral polarity utterances generated using 20 *wells* and 18 utterances from NXT with differing textual content.

### 6.4.3 Participants and Task

Participants were recruited through Prolific[9] and were selected to reside in the US, have English as a first language with no hearing impairments. We recruited 46 participants in total. Participants were presented with 78 stimuli, in a random order. On each trial, the audio of one stimulus and a rating scale were presented. As each stimulus was presented, participants were asked *What degree of agreement/disagreement is the speaker expressing?* The instructions stated that a rating of 1 represents *complete disagreement*, 99 represents *complete agreement* and 50 represents *neither agreement nor disagreement*. Unlike the

[8] Including the version of the carrier utterances synthesised with their ground-truth features.

[9] `https://www.prolific.com`

MOS instructions used in Chapter 3, here we are using a targetted evaluation which should assess the pragmatic effect that the prosody of the word *well* achieves, rather than a more nebulous term.

### 6.4.4 Statistical Testing

We treated the positive-polarity and negative-polarity utterances as separate datasets for statistical testing. We used linear mixed-effects models using the lme4 package (Bates et al., 2015) in R to test the effect of the four prosodic features on the rating of *agreement/disagreement*. We use the maximum random effects structure of the design which consists of a random effect for participant. Our dependent variable is the rating of agreement between 1-99 which we treat as continuous. Our fixed effects consist of the acoustic features which were used to cluster the *well* dataset (see Section 6.3.1), however because the coefficients are used to *condition* the model, and will show differences in output depending on neighbouring words, we extracted the acoustic features again from the synthesised stimuli.

```
agreement-rating ~ Legendre 0 + Legendre 1 + Legendre 2 +
                   duration + (1|listener)
```

## 6.5 Results

Two participants were removed, one for having issues with playing the audio and one for indicating in the post-question that they only focused on the lexical content of the stimuli. The results for the **positive-polarity** carrier utterances are found in Table 6.3. The mean rating for these stimuli was 75.50 (SD=14.58). We found that duration was a significant predictor of perceived agreement and had a negative relationship on the rating. As duration increases by one unit, the level of agreement perceived decreases. This might signal hesitation to the listener and seems to decrease the polarity of *yes*. We observe that when $F_0$ height (LC0) increases by one unit, the agreement increases by 0.62 and when the LC2 increases by one unit, the agreement rating also increases by 0.61. This suggests that these values modestly, but positively strengthen the assertion of *yes*.

Table 6.3: *Experimental Results per Polarity Type*

| | Fixed Effect | $\beta$ | St. Err | CI | p-value |
|---|---|---|---|---|---|
| Positive | LC0 | 0.62 | 0.25 | 0.13 - 1.11 | $< 0.05$ |
| | LC1 | -0.04 | 0.14 | -0.31 - 0.24 | $> 0.05$ |
| | LC2 | 0.61 | 0.24 | 0.14 - 1.09 | $< 0.05$ |
| | Duration | -7.13 | 2.79 | -12.59-1.67 | $< 0.05$ |
| Negative | LC0 | 0.75 | 0.35 | 0.06 - 1.44 | $< 0.05$ |
| | LC1 | -0.09 | 0.17 | -0.43 - 0.24 | $> 0.05$ |
| | LC2 | 0.72 | 0.35 | 0.04 - 1.40 | $< 0.05$ |
| | Duration | 14.35 | 2.79 | 8.89 - 19.81 | $< 0.01$ |

The results for the **negative-polarity** carrier utterances are in Table 6.3. The mean rating for the negative-polarity stimuli was 18.03 (SD=14.71). Similar to the results for positive-polarity, duration is a significant predictor of agreement, but has a *positive* relationship on agreement. As duration increases by one unit, the level of agreement perceived increases. This again suggests that an increase in duration shifts the polarity of the utterance, or weakens the strength of disagreement. Again, we observe that both the LC0 and LC2 exhibit a small, but positive relationship on the level of agreement. Thus unlike in the positive polarity condition, where these coefficients strengthen the assertion, in this case they weaken the negation. It should be noted that during statistical analysis, we observed heavy tails in the residual distributions. Though these models have been found to be robust to violations of the normality of residuals, the results warrant careful interpretation (Schielzeth et al., 2020).

## 6.6 Discussion

In this study, we presented a data-driven method for exploring prosodic patterns on the discourse marker *well* by clustering information about the duration and $F_0$ contour of unlabelled data. We used the centroids of the clusters to create synthetic stimuli for a listening test which tested the effect of the prosodic realisation of *well* on the degree of agreement expressed by the speaker. We found a consistent pattern regarding the effect of duration on both positive and negative short utterances. When *well* is realised with a longer duration, the rating tends away from disagreement for negative utterances, and away from agreement for positive utterances. This suggests a form of hedging for negative utterances and reluctant agreement (p 207 Aijmer (2015)) in the case of the positive utterance. Future work will aim to expand this study, by including more variation in the carrier utterances, to investigate whether this effect is consistent across different realisations of the textual and acoustic context following the discourse marker. As mentioned in the introduction, this case study serves to illustrate this method for a specific use-case, but is easily adapted to different words, contexts and research questions. For example, we could have asked participants how *certain* the speaker sounded, or any number of specific pragmatic functions.

There are a number of limitations of the study presented in this chapter. The first is in the data selection procedure. As mentioned, due to issues with the word-level forced alignment, we performed a second-pass of ASR on the spliced *wells*. We acknowledge, however, that ASR performance on isolated words, which have been spliced out of their original context, should be worse than in context. Because of this filtering step, we significantly reduced the quantity of data available for clustering. In this case, we chose a conservative approach. There was a clear trade-off between using fewer samples or using incorrectly-segmented samples. Because we were interested in single words, we chose the to verify the word alignments. The MFA model used in this study was trained on read speech, thus future work to address this issue could include fine-tuning the MFA model on a corpus, such as Switchboard Corpus I, which contains hand-corrected phone alignments. We expect that this would lead to a higher amount of data retention.

When clustering, we observed smaller, but more dense clusters in the periphery of the tSNE plot. Informal listening identified that these clusters contained quite salient prosodic features, which appeared to have a strong pragmatic meaning. In this study, however, we

did not perform qualitative analysis on each cluster. As mentioned in the introduction, this method can be used to identify less frequent but perceptually salient clusters in data (Ogden, 2012; Freeman et al., 2015a; Cruttenden, 1984). Future work could involve analysing particular clusters of realisations and the contexts in which these occur. Similarly, a second limitation of our clustering procedure is that we did not perform perceptual experiments on the samples in each cluster. To strengthen this method, future work should examine the perceptual similarity within and between clusters, especially clusters which are distant in the clustering space, in an experimental setting, rather than using informal listening by the authors.

A further limitation of this study is that we tested the effect of prosody on the perception of agreement and disagreement in *isolated* utterances using two carrier utterances. We chose these utterances because of their isolated meaning of negation or affirmation. This context-free interpretation was correctly identified as evidenced by the mean rating of both carrier texts. Though we did find an effect of three prosodic features on the stance expressed by a speaker, we cannot generalise to other contexts, nor can we say what the effect of these features would be given a preceding conversational context. Future work will investigate the effect of prosody on the stance expressed by the discourse marker by placing the utterances in a communicative context, as was done in Chapter 3. Again we reiterate that the method presented in this chapter allows for such adaptations.

We presented an example perceptual study to demonstrate how we can uncover pragmatic patterns in messy data, but this method could be applied to any other discourse marker. This method can be used by linguists, who want to create stimuli for experiments to test pragmatic theories, and by speech synthesis practitioners, who would like to explore the variation in their data. Using automatic tools to filter and label the data will not have perfect accuracy, but we hope nonetheless that found data can be used by speech technologists to do data exploration in the hope of finding new insights, and to generate new hypotheses. Exploring the data in this way is not meant to be strictly confirmatory of linguistic theory, and as such this work was largely exploratory in nature, and theory neutral. To apply this directly to phonetic research, in a more controlled approach, this method could be applied to an existing speech corpus which has been curated for the purpose of academic research, in particular one which contains some form of labelling and correct word-alignments. Finally, we clustered features related to $F_0$ and duration, but any number of features can be clustered if they can be extracted from the speech signal reliably, for example intensity or voice quality features such as HNR, jitter and shimmer. This method could therefore be used to identify utterances where certain clusters of features are likely to be found aiding exploratory analysis in phonetics.

## 6.7 Conclusion

As conversational systems move towards expressivity at a faster pace than our knowledge about pragmatics and conversational speech phenomena, it is important that we develop methods that can uncover these important patterns in data. In this chapter, we presented a methodology that enables discerning knowledge and hypotheses about pragmatic functions from large amounts of found data, and validates them with perceptual experiments. Though here we present a pilot study of how such a method can be used to test pragmatic functions of prosody in a data-driven manner with minimal labelling, there

are a number of improvements for future work, such as improving automatic tools for data filtering, and including formal cluster validation.

The method described in this chapter is not limited to the study of *well*, but can be used on any other word or phrase in conversational speech. Specifically, we are interested in how we can use the patterns found in the data to create stimuli, rather than using top-down knowledge, often based on studies of laboratory speech. We argue that this approach can be applied to both speech synthesis research and linguistic research. For speech synthesis research, this method would be particularly helpful for speech synthesis evaluation research as it would allow researchers to generate a wide number of stimuli from various dimensions of the acoustic space, or to identify patterns in the training data or in found data, which might be rare, but should be modelled. For linguistic research, larger quantities of data can be used to uncover patterns which may not be found in smaller datasets, such as those described in Section 6.2. Though found data might not be suitable for confirmatory hypothesis testing or corpus research, due the lack of control, it can allow researchers to generate new hypotheses, which can later be tested in more controlled environments.

# 7

# Investigating Turn-taking Prosody using Found Data and Speech Synthesis

This chapter is based on the following paper:

**O'Mahony, J.**, Lai, C., King, S. (2023) Synthesising turn-taking cues using natural conversational data. *Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW12)*, 75-80, doi:10.21437/SSW.2023-12

Author Contributions: First author – all work except for the prosodic analysis of the training data; Second author – prosodic analysis of training data, feedback on manuscript; Third author – editing and feedback on manuscript.

## 7.1 Introduction

As we saw in Chapter 6, one of the biggest challenges when working with large amounts of found conversational data is accounting for the variation in the data. To synthesise felicitous conversational prosody, we need to label aspects of the context that potentially exert influence on the prosodic form of an utterance. But as we saw in the previous chapter, we often do not know what these contextual features are. In the previous chapter, we therefore sought to *explore* the prosodic variation in the data using a bottom-up approach, i.e., by clustering prosodic features extracted from found data. Using a controllable TTS model, we synthesised renditions of the same text using the cluster centroid features to evaluate how each rendition affected the perceived stance of agreement/disagreement. Here TTS was used as a tool with which we can directly manipulate individual prosodic features in a controlled manner to assess their perceptual effect. But TTS can also be used as a tool to explore conversational prosody using a top-down approach, i.e., when we are fortunate to have a known contextual feature with which we can condition a model. By conditioning

a TTS model on a known aspect of context, we can synthesise utterances in each condition and assess whether the model has learned systematic prosodic correlates of each condition from the training data. In this chapter, we therefore illustrate how TTS can be used in an analysis-by-synthesis paradigm to explore the role of prosody in turn-taking.

As mentioned in Chapter 2, turn-taking is an important aspect of conversational interaction (Skantze, 2021) and describes the way in which speakers navigate the back-and-forth of conversation e.g. how they signal that they are giving up their speaking turn (turn-final) or whether they are holding the floor (turn-medial). Because of the fundamental importance of turn-taking in human interaction, this topic has received a large amount of attention in the fields of phonetics and psycholinguistics, and more recently there has been increased interest in predicting turn-taking events for dialogue systems (e.g. Edlund and Heldner (2005); Skantze (2021); Ekstedt and Skantze (2022b); Ekstedt et al. (2023)). Previous work has identified turn-taking cues that operate at various linguistic levels, including any combination of pragmatic, semantic, syntactic, or prosodic cues, all optional, but which have an additive effect (Hjalmarsson, 2011; Gravano and Hirschberg, 2011; Skantze, 2021).

The prosodic correlates of turn-taking events have been extensively studied in corpus research. Various prosodic features have been found to differ between turn-final and turn-medial utterances in the literature, including the presence of creaky voice (Heldner et al., 2019), longer turn-final IPUs (Gravano and Hirschberg, 2011), utterance-final lengthening in turn-medial IPUs (Hjalmarsson and Laskowski, 2011), and differences in speech rate (Brusco et al., 2017, 2020), as well as differing $F_0$ realisation (Brusco et al., 2020). However, there have also been conflicting results, as we will see in the Section 7.2. While the presence of systematic prosodic correlates in corpus studies suggest that prosody may play a role in turn taking, corpus research cannot directly inform us about the perceptual relevance of these cues (Gravano and Hirschberg, 2011). Therefore, in addition to corpus research, there has been a large body of research dedicated to studying the role of prosody in turn-taking from a perceptual standpoint.

Research studying the role of prosody in the perception of turn-transitions has been carried out using both offline and online experimental paradigms. Common paradigms in offline experiments include rating how likely a speaker is to give up their turn, using a Likert scale (Cutler and Pearson, 1985; Edlund and Heldner, 2005), and forced-choice tasks in which participants decide whether an utterance sounds turn-medial or turn-final (Cutler and Pearson, 1985), as well as pairwise comparisons of utterances (Cutler and Pearson, 1985; Zellers, 2017). Here we can see that offline experiments employ similar techniques to those used in speech synthesis evaluation, for example MOS tests and preference tests. To assess the implicit effect of prosody on the perception of turn transitions, online experiments have been proposed, such as the button-press task in which participants press a button in anticipation of a perceived turn-change (Ruiter et al., 2006; Bögels and Torreira, 2015). Again, there have been conflicting results with regards to the perceptual relevance of prosody.

One of the potential reasons for these conflicting results is that the approach to stimuli creation differs across studies. This is because creating natural, but controlled, stimuli for this task is difficult. Ideally, to independently assess the role of prosody, beyond other linguistic cues, the text and speaker should remain constant, differing only in prosodic realisation (Cutler and Pearson, 1985). For example, extracting utterances from real spontaneous speech (Gravano et al., 2016) does not afford us the ability to control textual

or other contextual factors across conditions (Gravano et al., 2011). To create controlled stimuli, with the same text spoken in different turn-contexts, some have used reenacted dialogue (Cutler and Pearson, 1985) or semi-scripted dialogue (Bögels and Torreira, 2015), but these stimuli might be void of the true turn-taking cues employed during spontaneous interaction. Similarly, prosodically manipulating existing spontaneous speech (Zellers, 2017; Ruiter et al., 2006) can lead to highly unnatural renditions. In Section 7.2, we present an overview of the experimental paradigms used in previous research, with a particular focus on stimuli creation.

So far we have seen that, both in corpus research and perceptual studies of turn-taking, there are potential disadvantages of the methods used. While providing invaluable insights into the prosody of turn-medial and turn-final utterances, corpus studies cannot measure perceptual significance of cues found (Brusco et al., 2020). In perceptual studies, the methods used to create stimuli may be sub-optimal if we want the stimuli to reflect the prosodic patterns found in actual conversational data. The issues mentioned above motivate the approach we take in this chapter, in which we employ a data-driven method to the creation of stimuli. By training a speech synthesis model using found conversational data, and additionally conditioning this model on turn-position, we can evaluate whether there are prosodic patterns which distinguish turn-final from turn-medial utterances and whether these patterns can be learned by the model. But additionally, by synthesising each condition and presenting a targeted offline evaluation of turn-taking, we can *also* assess whether these patterns lead to perceptible differences between turn-medial and turn-final conditions. As in the previous chapter, we use speech synthesis to aid our understanding of conversational speech by synthesising stimuli comprising of the same speaker and text in different contextual conditions.

Finally, the motivation for this work is not entirely linguistic in nature. As we have seen throughout this thesis, more and more research in speech synthesis is focusing contextualised conversational prosody (Guo et al., 2021; Cong et al., 2021; Mitsui et al., 2022; Li et al., 2022; Yamazaki et al., 2021). Ultimately we are interested in identifying contextual features which can explain variation in the speech signal, allowing meaningful control over prosodic renditions. Recall that in Chapter 4, we observed that our proposed speech synthesis model, which mixed found conversational data and read speech data, showed a slight degradation in performance for statements compared to the baseline model which was trained solely using read speech. Here we hypothesised that one of the reasons for this degradation was that the statements in the conversational data may have been either turn-medial or turn-final, meaning that the utterances may not have had consistent prosodic features. Thus, by studying the effect of turn-position, we can assess the benefit of adding this contextual factor to a TTS model. In this chapter we therefore ask the following research questions:

**RQ7.1** Does conditioning a TTS model on turn-taking enable the model to learn prosodic turn-taking cues from natural conversational speech that are perceptible to listeners?

**RQ7.2** Does conditioning a TTS model on turn-taking lead to increased turn-finality judgements over a baseline?

**RQ7.3** What prosodic cues distinguish natural turn-medial vs turn-final utterances and are they also present in synthetic speech?

To answer the above questions we:

- create a TTS training dataset using found spontaneous speech in which each utterance is characterised by its turn-position i.e. *turn-medial* or *turn-final*.
- use this dataset to train a speech synthesis model, which is additionally conditioned on turn-position, enabling this feature to be controlled at synthesis time.
- run an evaluation on IPU texts, taken from conversational data, to assess the perceived rating of turn-finality when no audio is presented. We then select the most ambiguous texts (*turn-ambiguous*) based on the ratings to be used in evaluation.
- synthesise renditions of each turn-ambiguous text in 1) the turn-final condition and 2) the turn-medial condition.
- evaluate the impact of each condition by performing a forced-choice listening test evaluating the perception of turn-finality between the turn-final and turn-medial rendition.
- perform prosodic analysis on the training data and synthesised renditions in both conditions.

## 7.2 Previous Work

### 7.2.1 Corpus Studies of Turn-taking Prosody

As mentioned, the prosodic correlates of turn-transitions have been extensively investigated in corpus studies. For example, Gravano and Hirschberg (2011) performed analysis of turn-medial and turn-final[1] inter-pausal units (IPUs) in the Columbia Games Corpus, a task-oriented dialogue between two speakers conversing to complete a game on a computer. Here, the goal was to investigate whether turn-medial and turn-final utterances could be detected using acoustic and linguistic cues derived automatically from the data. Additionally, they sought to evaluate the validity of the Duncan (1972) hypothesis stating that the more turn-yielding cues are present, the more likely a speaker-switch will occur. At the same time they evaluated the cues postulated by Duncan, including turn-final lengthening and differences in pitch realisation. They analysed differences between conditions, using both a symbolic representation of intonation (ToBI) and acoustic measurements of pitch, intensity and various voice quality features, such as jitter and shimmer. They found that, for turn-medial IPUs, a greater number of plateau contours occurred, while for turn-final IPUs, both high rise or falling pitch contours were common. For their objective acoustic measurements, they found that a positive $F_0$ slope was more common for turn-final IPUs, and that there was a lower mean $F_0$ and intensity in turn-final IPUs. Both turn-final and turn-medial IPU incurred IPU-final lengthening, but this was more pronounced in turn-medial IPUs, contrary to Duncan's hypothesis. Voice quality features, such as jitter and shimmer, were also a significant cue for turn-finality. Further, they found that cues had an additive effect – the more cues present, the higher the likelihood of a turn-switch.

[1] Gravano and Hirschberg (2011) also compared backchannel-inviting cues, but this is out of scope here.

Brusco et al. (2020) carried out a cross-linguistic study into the turn-taking cues employed by Argentine Spanish, Slovak and US English speakers, using Games corpora constructed similarly to that of Gravano and Hirschberg (2011). They used machine learning as a descriptive tool to assess the importance of static acoustic features, such as speech rate and IPU duration, and time-varying features, such as $F_0$, intensity and voice quality features. Examining the distribution of features in each condition, they found, similar to Gravano and Hirschberg (2011), that speech rate increased in turn-final utterances, again suggesting utterance-final lengthening in turn-medial position. Again, turn-medial IPUs were characterised by a flat pitch contour, while lower intensity was found in turn-final IPUs. Interestingly, similar cues were present cross-linguistically, a similar finding to Brusco et al. (2017). To assess the value of these features in turn-transition prediction, they created feature vectors per IPU consisting of the cues above. They trained Random Forest classifiers to predict whether IPUs were turn-medial or turn-final and assessed the importance of each feature in classification. They found that speech rate and IPU duration were the most important features for classification, followed by the final 200 ms of the pitch track and pitch slope. Jitter and shimmer showed little contribution, contrary to the study of Gravano and Hirschberg (2011), but HNR was found to be a distinguishing feature.

From the studies above, we can see that contrary to Duncan's hypothesis of turn-final lengthening, turn-medial utterances have been found to incur more lengthening. This result has been found by other studies including in Swedish (Hjalmarsson and Laskowski, 2011) and using kinematic data in US English (Purse and Krivokapić, 2023) and could be attributed to a reduction in "planning load" (Purse and Krivokapić, 2023, 1656) in turn-final utterances. Given the cross-linguistic and cross-speaker (Gravano and Hirschberg, 2011) tendency of this cue, we expect to see a similar tendency in the current study. Lower intensity in turn-final position has also been found to be a stable feature in discriminating turn-types (Gravano et al., 2011; Brusco et al., 2017; Wlodarczak and Heldner, 2022). Voice quality features, such as jitter and shimmer, show a mixed picture, between the studies of Gravano et al. (2011) and Brusco et al. (2020). However, there has been converging evidence of creaky voice and low periodicity in turn-final IPUs in more recent work (Heldner et al., 2019; Wlodarczak and Heldner, 2022). In the current study, however, we do not focus on voice quality features.

For $F_0$, there is converging evidence for flatter IPU-final $F_0$ contours in turn-medial IPUs, suggesting this might be a turn-holding cue (Gravano and Hirschberg, 2011; Brusco et al., 2017, 2020; Gravano and Vidal, 2014). For turn-final patterns there is more conflicting evidence across studies. For example, Brusco et al. (2017) found a rather uniform distribution of turn-final slope values, suggesting that all possible slopes can be present. Gravano and Hirschberg (2011) found evidence of both rising and falling contours, while Brusco et al. (2020) and Wlodarczak and Heldner (2022) found evidence of rising contours in turn-final position, while Zellers (2017) found no link between $F_0$ realisation and turn-transition type in a production study of Swedish. The variability found for turn-final intonation might be explained by other pragmatic effects. For example, the studies mentioned above do not distinguish between speech acts or dialogue moves, factors found to affect final rises (Lai, 2014). Further, the studies cited above used task-orientated dialogue, the data used in the present study is open-domain data, and $F_0$ features have been found to differ across genres (Lai, 2014). Thus, it is unclear, given the work discussed above, what the tendencies in $F_0$ realisation will be in our study.

Finally, as noted by Gravano and Hirschberg (2011) and Brusco et al. (2020), corpus studies can only provide information about statistical tendencies in the data, and not about the perception of these features. In this work we therefore motivate the use of TTS to explore turn-taking cues in found conversational data with the result being a perceptual stimulus with which we can conduct perceptual experiments. In the next section we explore previous work on the perception of turn-taking.

### 7.2.2 Experimental approaches to turn-taking

Corpus studies do not shed light on the perceptual relevance of prosody in turn-taking (Gravano et al., 2011; Brusco et al., 2020). Work into the perception of turn-transitions has therefore attempted to control or manipulate prosodic features across conditions to examine their effect on turn-taking perception. But creating controlled stimuli for such experiments is difficult. As noted by Gravano et al. (2011), one of the difficulties in corpus studies exploring turn-taking is that the data is unbalanced with respect to textual content and pragmatic context between turn conditions. For perception studies, this makes using naturally occurring speech, taken from corpora, similarly difficult. The *ideal* set of stimuli is one in which only the prosody differs and in which speaker, textual content and other factors remain the same (Cutler and Pearson, 1985). Nonetheless, speech taken from conversational corpora has been used in perceptual experiments, for example in Stephens and Beattie (1986); Gravano et al. (2016); Ruiter et al. (2006). The benefit of using naturally occurring speech is that it is the most ecologically valid.

Stephens and Beattie (1986), for example, presented listeners with turn-medial[2] or turn-final utterances from conversations in which the speakers agreed or disagreed with each other. In total they used 12 target speakers. Listeners were presented with the text or the audio, and in a forced choice paradigm decided whether the speaker was finished speaking or would continue to speak. When presented textually, participants were no better than chance in deciding whether an utterance was turn-final or turn-medial. When presented with audio, turn-medial utterances were identified 57% of time time, while turn-final utterances were no better than chance. There was a large amount of variation in turn-identification accuracy depending on the target speaker. The paradigm above can provide evidence that *something* in the speech signal affects identification of turn-transition. However, examining the accuracy in the text-only condition, we also observe that there was considerable variability in the correct identification of turn-medial and turn-final utterances, suggesting that some utterances contained textual cues of turn-transition type. Thus, given the variability between speakers in the audio condition and in the text condition, there are likely many confounds which may have affected the results. Without comparing the same text in different conditions, it is unclear what led to perceived differences.

To avoid the confound of stimulus text, Cutler and Pearson (1985) created stimuli using scripted dialogue. Cutler and Pearson (1985) developed small dialogue segments in which a turn-final or turn-medial utterance was embedded. The segments were identical, except that in the turn-medial case, another utterance followed. Following the elicitation of turn-medial and turn-final versions, they conducted a text-only perception test, to ensure that the utterances used in the experiments were maximally turn-ambiguous. After this,

[2]Note it was unclear in this study whether the utterances were bound by silences in the turn-medial case.

two perception tests were conducted. In the first test, isolated turn-medial or turn-final stimuli were presented, and listeners chose whether each stimulus sounded turn-medial or turn-final. In the second test, the turn-medial and turn-final version were presented side-by-side and listeners judged which stimulus sounded the most turn-final. In both experiments, participants did not score above chance in distinguishing turn-types. But this paradigm, though providing control over textual content, does not represent naturalistic conversational speech, as noted by the authors. This too may have affected results.

Other approaches use manipulated naturally-occurring speech to avoiding confounds of text and speaker. For example, Bögels and Torreira (2015) used semi-scripted dialogue to elicit target utterances which were short questions or long questions. Here the long questions textually matched the short question, but additionally were continued with a small number of words. To create the final stimuli for a button press task, segments from the long and short questions were spliced together in different combinations. However, splicing can also introduce artifacts, as was noted by the authors in later work (Bögels and Torreira, 2021). Other forms of stimulus manipulation include delexicalisation, achieved by low-pass filtering natural conversational data to remove lexical content. Here the motivation is that the delexicalised versions should only contain low frequency acoustic features to test whether prosody alone can impact turn-transition identification. In work employing this method (e.g. Ruiter et al. (2006); Gravano et al. (2016)), hearing the delexicalised renditions did not increase the accuracy in turn-transition identification. While this led Ruiter et al. (2006) to postulate that prosody was not important for turn-end perception, as noted by Gravano et al. (2016), this form of stimulus manipulation can remove durational and segmental features, which as we saw in the previous section are an important correlate found in corpus studies. Further, removing lexical content does not mirror actual human speech.

Other manipulations which have been proposed include adapting the $F_0$ of a stimulus in a step-wise manner or adapting duration of individual segments. For example, Zellers (2017) used natural recordings of Swedish task-oriented dialogue, and selected utterances which were syntactically complete. Utterances were adapted to have differing $F_0$ and duration of the final unstressed syllable using TD-PSOLA. In a forced-choice paradigm, two renditions were presented and listeners chose which sounded most turn-final in one condition, and which sounded turn-medial in the other condition. She found for Swedish, that duration of the final unstressed syllable was the most reliable cue, with longer items being perceived as turn-medial. $F_0$ height was also significant but only when the duration of the syllable was the same, with higher $F_0$ indicating turn-medial. Again she found an additive effect, more cues led to more correct responses. As we discussed in Chapter 5, this form of manipulation can lead to unwanted acoustic artifacts and manipulating speech using top-down information may not reflect how speech is actually realised in conversation.

Finally, in more recent work, Lameris et al. (2024) used TTS to explore the role of creaky voice in turn-transition perception. First, a corpus of conversational speech was annotated automatically for the presence of creak. This corpus was then used to train a Tacotron (Shen et al., 2018) TTS model, conditioned on word-level *creak percentage* (Lameris et al., 2024). Following this, different renditions of the same texts were synthesised. Participants were asked "How likely is it the speaker has finished speaking?" (Lameris et al., 2024, p.16061) on a scale of one to seven. Results showed that the presence of creak led to higher ratings of turn-finality, suggesting that voice quality is a potential turn-taking cue.

In summary, in this section we have seen that many different approaches to stimuli

creation have been proposed to study the role of prosody in turn-taking. Overall, there is a significant trade-off between control over text and speaker, versus the external validity of the stimuli. However, we have also seen, as shown by the work in Lameris et al. (2024), that as speech synthesis models improve, we can synthesise conversational speech in which we can both control the speaker and textual content while manipulating individual cues, similar to our work presented in Chapter 5 and 6. In this chapter, however, we take a different approach: we use speech synthesis to both learn prosodic correlates across turn-transition conditions in a corpus of natural speech and to test whether these prosodic correlates are perceptible by speakers.

## 7.3 Method

### 7.3.1 Data

#### Conversational Data

We use the CANDOR Corpus (Reece et al., 2023), introduced in Chapter 6, which is comprised of 1656 open-domain online conversations between two speakers, recorded in separate channels, and transcribed automatically. We first split the data into IPUs using a silence threshold of 200 ms. IPUs were removed if they comprised only of backchannels to reduce the chance of overlapping speech (out of scope here). To characterise the position of each IPU in a turn, we implemented communicative state classification (Heldner and Edlund, 2010). We selected target IPUs if no overlap occurred on either side of the IPU, leaving 262446 [left-context]-[target]-[right-context] triplets, labelled according to who spoke in the right-context IPU, i.e., *same* (turn-medial) vs *different* (turn-final) speaker to the target. An example of this selection strategy is seen in Figure 7.1. Here we can see that for Speaker A, IPU A1 contains no adjacent overlaps and is turn-medial because the speaker continues to hold the floor. Similarly IPU A2 is also chosen, but is turn-final because Speaker B takes the floor. IPU A3, contains overlap, and is therefore not chosen.

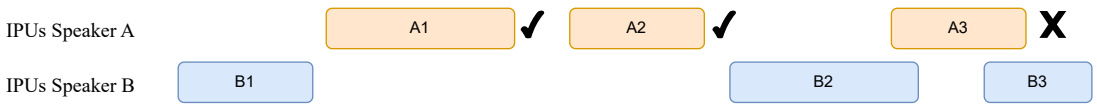


Figure 7.1: *Depiction of IPU Selection*

For initial data selection we chose target IPUs with duration 1-11 s due to memory constraints for TTS model training, and only IPUs surrounded by no more than 2 s of silence on either side. All target IPU texts containing symbols, numbers or acronyms were removed to avoid text normalisation issues. We removed targets containing a question-mark as a proxy for questions because our focus is declarative prosody. We then stripped all automatically-inserted punctuation. We calculated speech rate using canonical syllables per second and only retained target IPUs with speech rate of 2-6 syllables per second. We finally removed triplets where the left-context or target IPU had fewer than three words.

We then made two subsets of data. The first (modelling) consisted of all speakers with more than 10 minutes of speech, to be used for modelling. The second (heldout) contained

speakers with less than 10 minutes of speech, to be used as a source of text for our evaluation, and never used in training. After initial filtering we had 51093 targets (turn-medial 40739; turn-final 10354) totalling 56.72 hrs. We aligned the modelling data using the MFA (McAuliffe et al., 2017) and removed all targets where alignment failed, and all speakers where more than 5% of their utterances failed during forced alignment.

### Read Speech Data

Initial tests showed that training models on only the above data was not feasible, due to limited data quantity per speaker and variable recording conditions. Therefore, following previous work on synthesising spontaneous speech using found data (Szekely et al., 2019), we used the LJ Speech dataset to pre-train our model to enable transfer learning. In addition, we chunked the utterances in the dataset at punctuation to shorten the utterances to make them more comparable to IPUs in the conversational data[3]. We removed punctuation in the read speech data before training to ensure that it mirrored the conversational data. We removed utterances with fewer than 3 words, leaving 16000 utterances for training and 100 for development.

### Final Data Selection

Natural conversations often exhibit class imbalance – there are often far more turn-medial IPUs – so further selection was required to improve class balance. We first took all data from the 41 speakers each with 15-20 minutes of speech to ensure that we had enough speakers with sufficient data. To this, we added all data from speakers with a high number of turn-final IPUs, plus all turn-final IPUs from the remaining speakers.

We then calculated the number of turn-final IPUs per speaker and, by random selection, ensured the data contained the same number of turn-medial IPUs for that speaker. Therefore, for each speaker in the dataset, both turn-final and turn-medial conditions were seen. The resulting dataset contains all available turn-final IPUs, but still has some class imbalance: Table 7.1. We then partitioned the dataset into 24000 utterances (23914 CANDOR + 86 LJSpeech) for training, 100 for development, and the remaining 242 for testing. There are 212 unique speakers in the final dataset (including LJ).

### Training Data Acoustic Analysis

$F_0$ and intensity contours were extracted using Praat at 10 $\mathrm{ms}$ intervals. $F_0$ parameter settings were automatically determined (Evanini and Lai, 2010), with per speaker pitch floor and ceiling based on global $F_0$ values. $F_0$ and intensity were then normalised to make values comparable across speakers and samples. Intensity was normalised by subtracting speaker mean per IPU. F0 was converted to semitones relative to speaker global mean in Hz. Values more than 2.5 standard deviations from the utterance mean were removed as outliers. We also checked for octave jumps ending outside of the 5th and 95th quantiles per IPU and removed values outside of those. Finally, we repeat the process if octave jumps are still detected, with the F0 floor and ceiling determined by the current utterance F0 values, i.e,. using a narrower F0 range.

We characterised $F_0$ and intensity contours using using Legendre Polynomial (LP) decomposition, which we described in Chapter 4. Recall the first 3 LP coefficients

[3]This new utterance segmentation was provided by Niamh Corkey

represent $F_0$ height, slope, and convexity of the contour respectively. Coefficients were determined using least squares fit of an order 5 Legendre series over a specified interval, time normalised to span [-1,1]. To analyse potential differences in turn-medial and turn-final prosody, we calculate LP coefficients for $F_0$ and intensity over the last 500 ms of the IPUs in the training data, inspecting only the first 3 coefficients. We also calculate the speech rate (syllables/second) over the whole IPU based on phone alignments.

Table 7.1: *Descriptive statistics and median acoustic feature values for final conversational corpus.*

| | Turn-Medial | Turn-Final |
|---|---|---|
| Total Turns | 15788 | 8468 |
| Duration (hrs) | 17.99 | 8.74 |
| Mean Tokens | 13.00 | 12.03 |
| Female Utterances | 5740 | 3002 |
| Male Utterances | 10048 | 5466 |
| Acoustic Features | | |
| F0 height (LP coeff 1) | -1.05 | -1.12 |
| F0 slope (LP coeff 2) | -0.40 | -0.15 |
| F0 convexity (LP coeff 3) | 0.13 | 0.18 |
| Intensity height (LP coeff 1) | -0.59 | -0.76 |
| Intensity slope (LP coeff 2) | -2.56 | -3.50 |
| Intensity convexity (LP coeff 3) | -2.44 | -2.99 |
| Speech rate (syll/s) | 4.00 | 4.10 |

The median values of the features above are shown in Table 7.1 for both turn-medial and turn-final IPUs. We found significant differences in the first 3 LP coefficients for $F_0$ and intensity, as well as speaking rate (Wilcoxon ranked sum test, $p < 0.05$). Overall, turn-final IPU ends are characterised by a lower, flatter F0 contour and lower overall intensity. We observe a slightly faster speaking rate in the turn-final condition, but differences between conditions are small. The results regarding intensity and speech rate closely mirror those of corpus studies discussed in Section 7.2.1, though we find a flatter slope in turn-final position, contrary to the corpus studies presented, which often found flatter $F_0$ contours on turn-medial utterances.

### 7.3.2 Model

To distinguish turn-position in a speech synthesis model, we trained a **TURN** and a **BASELINE** FastPitch (Łańcucki, 2021) model with identical architectures. We used an adapted version of the FastPitch model which uses phone durations derived from MFA textgrids as input to the duration predictor. These phone durations are also used to upsample each phone to the correct number of frames during training before the prediction of the mel spectrogram which in turn controls the duration of the final utterance. The automatically aligned phone sequence from MFA was also used as our input to the model (including silence tokens). We applied global mean/variance per-speaker F0 normalisation. Turn-conditioning is incorporated using an embedding

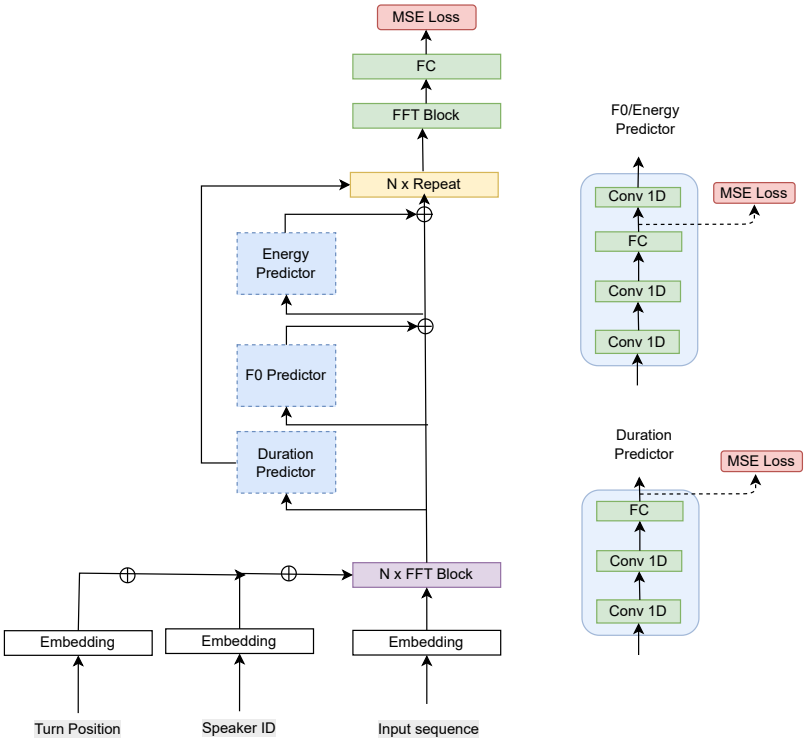


Figure 7.2: *Model architecture of FastPitch conditioned on turn-taking status (adapted from (Łańcucki, 2021)*

table, whose output is summed to the encoder output and speaker embedding, before being passed forward to the variance adapters (see Figure 7.2). The turn-condition input has three possible values: 0 for baseline; 1 for turn-medial; 2 for turn-final.

The only difference between **BASELINE** and **TURN** is the value of the turn-condition input. For **BASELINE**, it is fixed to 0 throughout all training and during inference. For **TURN**, it is set to 0 for read-speech data, to 1 for conversational IPUs labelled as turn-medial, and to 2 for conversational IPUs labelled as turn-final.

Pre-training was identical for both models, using only the read speech data with a batch size of 16 for a total of 200k steps. For both **BASELINE** and **TURN**, we completed training with a batch size of 32 for an additional 525k steps using the conversational data described above, plus 86 LJSpeech utterances. For both models the speaker embedding table contained 500 speaker codes and the turn-condition embedding table contains 3 codes (0 for baseline; 1 for turn-medial; 2 for turn-final). Each model was trained on on a single NVIDIA GeForce GTX 1080 Ti GPU. After FastPitch inference, waveforms were generated using the HiFi-GAN universal vocoder (Kong et al., 2020) with a denoising factor of 0.01.

## 7.4 Subjective Evaluation

As we saw in Chapter 2, most previous studies evaluate context-aware models using standard testing procedures like Mean Opinion Score, often testing for an increase in *naturalness* or *appropriateness*. But as we saw in Chapter 3, a more targeted evaluation is often warranted to assess contextual effects. So instead, we use a method inspired by the work of Cutler and Pearson (1985). However, we use conversational speech synthesis to generate stimuli, rather than scripted dialogue.

### 7.4.1 Test Materials

#### Text Selection

Turn-taking cues are not only prosodic. In fact, the textual content is a strong cue, with prosody having an *additive* effect (Skantze, 2021; Hjalmarsson, 2011). Because of this, when evaluating prosodic turn-taking cues it is important to choose turn-ambiguous texts, where the likelihood of a turn-end cannot easily be judged using text alone (Cutler and Pearson, 1985).

We selected 600 random utterance texts from our held out data, 300 turn-final and 300 turn-medial, which were then filtered for personal identifying material, controversial topics and profanity, leaving 532 sentences. We divided these into 4 groups of 133. Each group of texts was presented to 10 participants who were asked to rate each sentence, presented as text only, for turn-finality on a scale of 1-5 where a rating of 3 indicates maximum uncertainty that the speaker had finished talking.

On the initial instruction page of the experiment, participants were briefed on the task with the general instruction being: *For each utterance text, we would like you to indicate how likely it is that the speaker has finished taking their speaking turn in the conversation.* We used the rating scale wording from (Cutler and Pearson, 1985), where "where 1 represented *definitely still has more to say*, 2 *probably still has more to say*, 3 *could be going on or could be finished*, 4 *probably finished* and 5 *definitely finished*." (Cutler and Pearson, 1985, p.145).

We then took the median and mode rating of each text and chose utterances with only one mode, a mode of 3 and a median rating between 2.75-3.25.

**Target Speaker Selection**

As expected, given the large number of speakers with variable data quantity and recording conditions, the models could not synthesise all speakers with good quality. Expert listening (by the first author) to synthetic speech from the **BASELINE** model for the 53 speakers with more than 10 minutes of training speech was used to eliminate potential target speakers before the formal listening test, reducing the pool to 24 speakers. Speakers were chosen based on the overall impression of naturalness, mostly pertaining to the intelligibility of the speech output and the how well the speaker could be vocoded. For some speakers there were more noticeable vocoder artifacts than for others. We then conducted a pre-test to select speakers for further evaluation of turn-taking cues. 20 participants then rated the naturalness of 5 synthetic utterances for each speaker on a scale of 1-5 (total utterances = 120), presented as speech only, in a randomised order. Using the mean ratings per speaker, we picked the best 5 target speakers (Table 7.2) for use in the following listening tests.

Table 7.2: *Target Speaker Corpus Training Information*

| Speaker | Total Utts | Turn-medial | Turn-final | Turn-final % | Mean Naturalness Rating Pretest |
|---|---|---|---|---|---|
| 200 | 245 | 200 | 45 | 17.71% | 2.85 |
| 176 | 184 | 168 | 16 | 8.70% | 2.93 |
| 156 | 273 | 190 | 83 | 30.40% | 2.77 |
| 143 | 238 | 216 | 22 | 9.24% | 2.82 |
| 48 | 192 | 174 | 18 | 9.38% | 2.91 |

### 7.4.2 Participants

In all of the experiments, listeners were recruited using Prolific[4] and reported being English native speakers, residing in the US, with no hearing impairments. At the beginning of the experiment, we asked participants if they were using headphones and at the end whether they could play all of the audio.

### 7.4.3 Statistical Analysis

For all of the experiments below we used binomial mixed-effects regression models with a logit-link function (Bates et al., 2015) due to lack of independence in listeners and stimuli (as discussed in previous chapters). We included stimulus and listener as random effects. No predictors were included, making it a mixed-effects equivalent to an exact binomial test with the null hypothesis being that participant choice does not differ from chance. For each experiment, *choice* denotes the chosen audio for most final-sounding and the model is specified as:

```
choice ~ 1 + (1|listener) + (1|stimulus)
```

[4]https://www.prolific.co

### 7.4.4 Experiment 1 – Finality Judgements, Turn Model

To answer question RQ7.1 (end of Section 7.1), we tested whether conditioning the **TURN** model on the turn-medial vs. turn-final flag led to a perceptible difference in turn-finality for listeners. If turn-finality is shown to increase in the turn-final condition, it would suggest that the data does include cues for turn-finality in that condition and that the model has learned these prosodic correlates. For each of the five target speakers, we synthesised two versions of the 50 turn-ambiguous[5] texts using the **TURN** model, one turn-medial, the other turn-final. We then conducted a separate listening test for each speaker. In each test, 20 listeners were presented with 50 pairs of synthetic utterances. The order was shuffled and the within-pair order randomised, per listener. We asked listeners *Which of the following sounds like the speaker is finished talking?*.

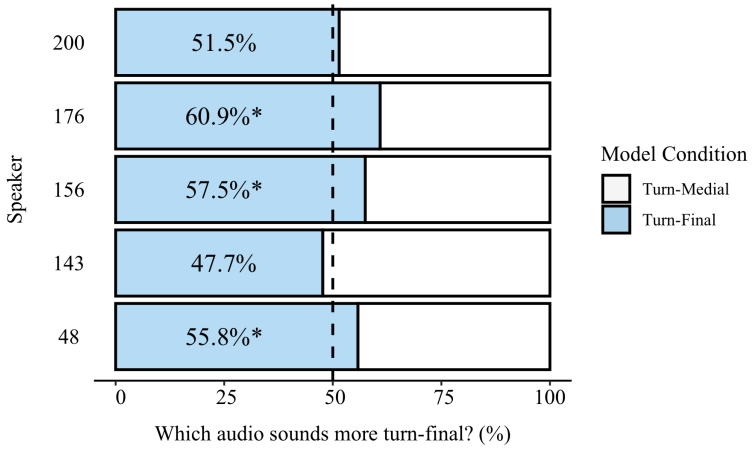


Figure 7.3: *Results for Experiment 1 comparing **TURN** generating turn-medial vs **TURN** generating turn-final per speaker*

20 participants took part in each listening test (total=100; of which we removed 3 for not wearing headphones and 3 for having issues playing audio). We analysed each listening test (i.e., each target speaker) independently and report the results in Table 7.3 and Figure 7.3. We found a significant majority choice for the turn-final synthesis for 3 speakers, with 2 speakers showing no significant difference in choice. This demonstrates that our model is able to learn patterns of turn-taking, but that the results are speaker-dependent.

### 7.4.5 Experiment 2 – Finality Judgements, Turn Model vs Baseline

To answer Question RQ7.2 (end of Section 7.1), we evaluated whether our **TURN** model in turn-final condition sounds more turn-final than **BASELINE**, which is trained on both turn-final and turn-medial IPUs without indicating their turn position. We synthesised the same 50 turn-ambiguous texts used in Experiment 1, but this time we compared the output of the **BASELINE** model with the **TURN** model operating with the turn-condition input set to turn-final. Again, we tested each speaker in a separate listening test, using the same design as Experiment 1.

[5] Samples: https://johannahom.github.io/SSW-2023/

Table 7.3: *Results from the linear mixed-effects model of Experiment 1 comparing **TURN** generating turn-medial vs **TURN** generating turn-final, per speaker*

| Speaker | $\beta$ | Probability Estimate | Confidence Interval | p-value |
|---|---|---|---|---|
| 200 | 0.06 | 0.52 | 0.46 - 0.57 | $> 0.05$ |
| 176 | 0.49 | 0.62 | 0.56 - 0.68 | $< 0.05$ |
| 156 | 0.33 | 0.58 | 0.52 - 0.64 | $< 0.05$ |
| 143 | -0.11 | 0.47 | 0.40 - 0.54 | $> 0.05$ |
| 48 | 0.25 | 0.56 | 0.51 - 0.61 | $< 0.05$ |

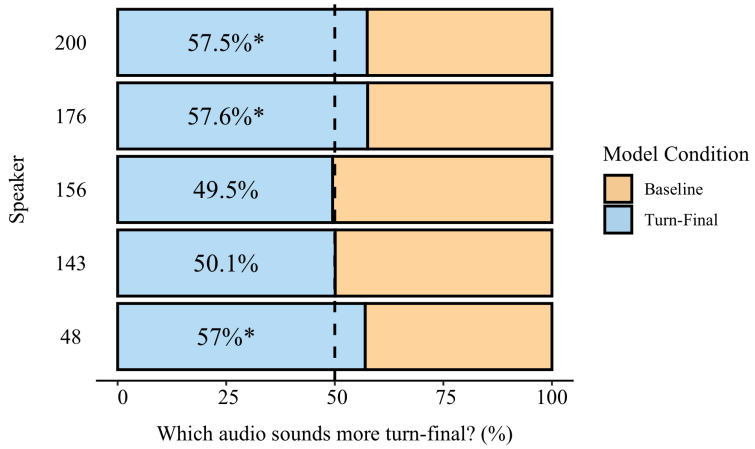


Figure 7.4: *Results of Experiment 2 comparing **TURN** generating turn-final vs **BASELINE** per speaker*

20 participants took part in each listening test (total=100; of which we removed 5 for not wearing headphones and 4 for having audio issues). The results are summarised in Table 7.4. Here we can see that speaker 176 and 48 mirror Experiment 1: the turn-final condition sounds more turn-final than baseline. Speaker 143 remains on par with baseline. We see a change for speaker 200 who sounded more turn-final than baseline here, but who had no significant preference to turn-medial in Experiment 1. Speaker 156 on the other hand had a significant difference in Experiment 1, but now is no different to baseline; this might be due to this speaker having a higher number of turn-final IPUs in the training data (Table 7.2), making **BASELINE** sound more turn-final. Overall, the results taken together indicate that **TURN** is able to produce turn-taking cues more effectively than **BASELINE**, for most, but not all speakers.

Table 7.4: *Results from the linear mixed-effects model of Experiment 2 comparing **TURN** generating turn-final vs **BASELINE**, per speaker.*

| Speaker | $\beta$ | Probability Estimate | Confidence Interval | p-value |
|---|---|---|---|---|
| 200 | 0.31 | 0.58 | 0.53 - 0.63 | $< 0.05$ |
| 176 | 0.33 | 0.58 | 0.52 - 0.64 | $< 0.05$ |
| 156 | -0.03 | 0.49 | 0.42 - 0.56 | $> 0.05$ |
| 143 | 0.004 | 0.50 | 0.45 - 0.55 | $> 0.05$ |
| 48 | 0.34 | 0.58 | 0.50 - 0.66 | $< 0.05$ |

## 7.5 Objective Evaluation

To investigate the cause of the speaker-dependent results from Experiments 1 and 2, and to answer question RQ7.3 (end of Section 7.1), we extracted the same acoustic information as for the training data. By analysing the acoustic output of each model condition along with the training data for each speaker, we hope to provide insight into which cues may be important for listeners when determining turn-finality. We are also interested in whether cues found in the literature are also found in our turn-final and turn-medial synthesised renditions. We hypothesise that speakers who show more differences in cues between conditions and should show large preferences in turn-finality ratings as cues should have an additive effect (Hjalmarsson, 2011; Skantze, 2021; Zellers, 2017; Gravano et al., 2011). Measuring cues and comparing these cues to those found in the training data of each speaker can also provide insight into how much information is learned across speakers in the data and how much is constrained by speaker conditioning.

### 7.5.1 Comparison of prosodic features in turn-final and turn-medial IPUs: *natural speech*

Table 7.5 shows differences in acoustic features for turn-medial and turn-final IPUs for our target speakers in the training corpus. Compared to the training data on aggregate, we observe speaker-specific correlates for turn-taking, but not all features show significant differences. Moreover, differences are not always in the same direction. For example, Figure 7.5 shows $F_0$ height (1st LP coefficient) across conditions for target speakers. We see

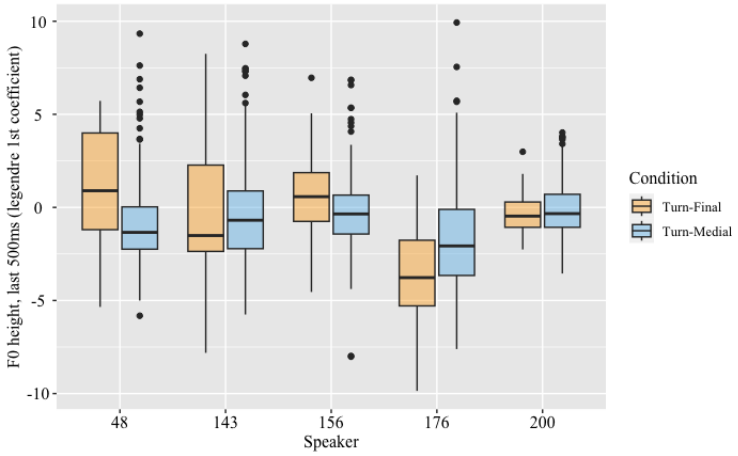


Figure 7.5: *F0 height per speaker, Last 500 ms*

that Speaker 176 has a lower $F_0$ for turn-final IPUs, while speaker 48 has higher F0. In fact, listening to samples suggests that speaker 48 has utterance final pitch rises as a default, i.e. uptalk. The variation in $F_0$ height, as well as the presence of uptalk in one of our speakers, highlights the variability of this feature across speakers as was shown in the studies in Section 7.2.1. Speaker 48 also exhibited higher intensity (height) in the turn-final condition (though not significantly so), while the 4 other speakers showed lower intensity turn-finally. Mean speech rate was slower at the end of turn-medial IPUs for two speakers, however the differences were not significant for any of the speakers. Given the data imbalance between turn-medial and turn-final IPUs per speaker, however, statistical testing may not be reliable.

Table 7.5: *Significant differences ($p < 0.05$) between turn-medial and turn-final IPUs for speakers (Wilcoxon ranked sum test) in natural speech*

| | 48 | 143 | 156 | 176 | 200 |
|---|---|---|---|---|---|
| F0 height | ✓ | . | ✓ | ✓ | . |
| F0 slope | . | . | . | . | . |
| F0 convexity | . | ✓ | . | ✓ | . |
| Intensity height | . | . | ✓ | . | . |
| Intensity slope | . | ✓ | ✓ | . | . |
| Intensity convexity | . | . | . | ✓ | . |
| Speech rate (syll/s) | . | . | . | . | . |

### 7.5.2 Comparison of prosodic features in turn-final and turn-medial utterances: *synthesised speech*

First for Experiment 1, we compare the values of various prosodic correlates found in Table 7.6 between the turn-medial and turn-final output per speaker. As we can see, across all speakers, we find a significant difference between turn-medial and turn-final speech rate and final word duration. This mirrors results found in the study of Brusco et al. (2020) and suggests that these features may have been learned by the model for the training data on

aggregate, because speaking rate was not shown to differ significantly between the speakers' natural turn-types (see previous section). In Figure A.4 in the appendix, we can see the direction of this difference, with all speakers showing an decrease in final word duration when synthesised as turn-final and speech rate being significantly faster than in turn-medial position, again this mirrors findings in previous corpus research. Interestingly, differences in these features do not directly lead to an increase in turn-final choices in Experiment 1, as in the case of speaker 143 and 200. We found however, and if we look at the average pitch height of the final word (Figure A.4 in the appendix) that speaker 143 shows higher $F_0$ in turn-final position which may have led to more turn-medial judgements. Similarly, speaker 200 compared to the other speakers shows no significant difference between turn-final and turn-medial conditions for $F_0$ height.

Table 7.6: *Significant differences ($p < 0.05$) between turn-medial and turn-final synthetic output for each speaker (Wilcoxon signed-ranks test)*

| Feature | 48 | 143 | 156 | 176 | 200 |
|---|---|---|---|---|---|
| Global F0 height | ✓ | ✓ | . | ✓ | . |
| Global F0 slope | . | . | . | . | ✓ |
| Global F0 convexity | . | . | . | . | . |
| Final word F0 height | . | ✓ | ✓ | ✓ | . |
| Final word F0 slope | . | . | . | . | . |
| Final word F0 convexity | . | . | . | . | . |
| Intensity height | . | . | . | . | . |
| Intensity slope | . | . | . | . | . |
| Intensity convexity | . | ✓ | . | . | . |
| Final word log duration | ✓ | ✓ | ✓ | ✓ | ✓ |
| Speech rate (syll/s) | ✓ | ✓ | ✓ | ✓ | ✓ |

### 7.5.3 Comparison of prosodic features in turn-final and baseline utterances: *synthesised speech*

For the stimuli in Experiment 2, we do the same comparison, comparing the output of the turn-final condition and the baseline model output. Compared to the results in Table 7.6 in Table 7.7 we see fewer significant differences between features in the baseline and turn-final condition. This suggests that our **TURN** model leads to starker differences between the turn-medial condition compared with the baseline along these dimensions. Interestingly, the speakers which show significant differences in turn-finality judgments in Experiment 2 all show a significant difference in speech rate.

## 7.6 Discussion

In this study, we found that conditioning a model on IPU turn-position leads to increased perception of turn-finality compared to the turn-medial case and the baseline, but that this is speaker-specific. Specifically, speaker-specific differences might arise from differences in speakers' natural prosodic cues, but might also be impacted by the quantity of each

Table 7.7: *Significant differences ($p < 0.05$) between BASELINE and TURN turn-final condition for speakers (Wilcoxon signed-ranks test)*

| Feature | 48 | 143 | 156 | 176 | 200 |
|---|---|---|---|---|---|
| Global F0 height | . | . | . | ✓ | . |
| Global F0 slope | . | . | . | . | . |
| Global F0 convexity | . | . | . | . | . |
| Final word F0 height | . | ✓ | ✓ | ✓ | ✓ |
| Final word F0 slope | . | . | . | . | . |
| Final word F0 convexity | . | . | . | . | . |
| Intensity height | . | ✓ | . | . | . |
| Intensity slope | . | . | . | . | . |
| Intensity convexity | . | . | . | . | . |
| Final word log duration | . | ✓ | . | ✓ | ✓ |
| Speech rate (syll/s) | ✓ | . | . | ✓ | ✓ |

turn-type seen in training. For example, speaker 156 learns to sound more turn-final than the turn-medial condition in the **TURN** model, but sounds equally turn-final compared to the baseline. This is potentially due to this speaker having more turn-final IPUs than other speakers in the training data leading to the baseline renditions sounding more ambiguous. To test this, future work should statistically compare the acoustic features of the turn-medial synthetic speech and the baseline synthetic speech. Future work should also adapt the experiment instructions to ask who is more likely to continue which will help to gain more insight into how instructions might affect results (Zellers, 2017).

Though the number of IPUs in each turn-position per speaker might affect the comparison between the turn-model and the baseline, we see that for speakers with very few turn-final IPUs in training, a benefit was found from training on a large amount of turn-final IPUs across other speakers. For example, speakers 48 and 176 showed a significant increase in turn-final judgements in the **TURN** model conditioned with the *turn-final* tag compared to renditions synthesised with the *turn-medial* tag, even though the percentage of training instances in turn-final position was under 10% for each speaker. The effect of learning across speakers can also be seen when we compare the differences in features between the turn-final and turn-medial IPUs in the training data, and the differences in the synthetic speech between the turn-medial and turn-final condition. In the synthetic speech, all of our target speakers exhibited final-word lengthening in turn-medial IPUs and an increased speech rate in the turn-final condition, mirroring results found in corpus studies. This suggests that this was learned corpus-wide as these features were not found in their natural productions.

However, the benefit of learning across speakers was not present for speaker 143. Speaker 143 did not show an increase turn-final perception, when comparing the renditions of turn-medial and turn-final synthetic speech, although some acoustic correlates were learned as evidenced by the objective evaluation. Thus there may also be interactions with the speaker specific features learned by the model in the speaker embedding. Such speaker-specific differences, however, were also found in the perception of turn-finality in the study of Stephens and Beattie (1986). To improve this paradigm, it is

therefore important that other sources of systematic variation are accounted for between speakers, especially if these sources of variation can account for differing turn-taking behaviour. For example, we did not have access to dialectal information in the CANDOR corpus. Future work should seek to account for such sources of variation by having data labelled or by employing the use of dialect recognition models. In this way, we could condition the model, not only on the turn-taking label, but on the turn-taking label given a specific dialect.

Initial results suggest that speech rate and word-final lengthening differences between utterances might be a helpful cue, but possibly only in combination with other factors such as a lowering in $F_0$. Again these cues are most likely additive (Hjalmarsson, 2011) and interact with each other (Ekstedt and Skantze, 2022a). We also found that the direction of $F_0$ height was not the same across speakers and across conditions. This was particularly true for the turn-final synthesis of speaker 143 who had a higher $F_0$ level on the final word compared to the turn-medial synthesis. This may have lead to different turn-finality interpretations. A future direction would be to analyse the turn-finality ratings per experiment per stimulus and correlate these ratings with acoustic cues found in the synthesis to gain more insight into which specific cues and combinations might be helpful to listeners in judging turn-finality. This might be fruitful as we found differences in ratings across utterances, even for speakers who showed no significant differences overall between conditions. Future work should also perceptually evaluate turn-taking cues found in the ground-truth recordings of the target speakers to create an baseline with natural speech. Finally, other work presented in Section 7.2.1 investigated voice quality features, such as the presence of creaky voice and correlates hereof such as jitter and shimmer and HNR, and future work should aim to investigate these features. It is currently unclear whether speakers in our dataset use voice quality as a turn-taking cue, whether our model produced such cues in different turn positions, and whether our vocoder could successfully synthesise different voice quality cues. Voice quality could be investigated in future work by analysing voice quality features, such as HNR, in the training data and synthesis output, or by detecting creaky voice in the training data automatically similar to Lameris et al. (2024).

Given the results above, we now turn to discussing the advantages and disadvantages of this method, both from a speech synthesis perspective and from the linguistic perspective. Recent work by Ekstedt et al. (2023), published after this study, investigated turn-taking features in both commercial TTS models, and in a Tacotron TTS model trained using conversational data taken from podcasts. They found that the model trained on podcast data was able to reliably signal turn-holding cues, as measured using the Voice Activity Projection model (Ekstedt and Skantze, 2022b), which predicts the likelihood of turn-transition types. For turn-final signals, the commercial systems significantly outperformed the podcast model, suggesting they produce adequate turn-final prosody. Given the results in our study, this result may be unsurprising. As we saw, in conversational speech, when segmented into IPUs, or breath groups – as in the case of the podcast data in the study of Ekstedt et al. (2023) potentially using breath groups – there is often a significantly higher number of turn-medial utterances. This would explain why the podcast model far out-performed the commercial systems when producing turn-medial prosody. Due to this class imbalance, it is conceivable that turn-final cues are not reliably learned by a TTS model without balancing the turn-position classes and additionally conditioning on this feature. For the commercial systems, it is conceivable that the training data has been recorded specifically for dialogue and in particular for rapid back-and-forth exchange,

which is common in voice assistants. Though, we should note that without having details of the training data used in commercial systems, this remains a conjecture. Thus from a speech synthesis perspective, the results in the work suggest that taking turn-position into account is valuable, as evidenced by the increase in turn-finality judgments across the majority of speakers in our study.

From a linguistic perspective this method provides an elegant way to explore turn-taking cues in natural speech. Unlike corpus studies, which provide information on acoustic correlates across turn-position types, this method allows us to use TTS to learn the same features, but the result here is a perceptual stimulus. Though an increase in turn-finality judgements was not found in all speakers, as discussed above, it is clear that specific patterns were learned across speakers and that these patterns led to an increase in turn-final judgements in the majority of our target speakers. Moreover some of these patterns were similar to those found in previous corpus studies, such as turn-final word lengthening and increased speech rate. As discussed above, future work could explore the inclusion of dialectal information in the speech synthesis model to explore the realisation and perception of turn-taking cues from a cross-dialectal and sociolinguistic perspective.

However, there are a number of limitations to this method. First, given the mixed results in our objective evaluation, we still do not know exactly which cues led to the increase in turn-finality perception. Again, further statistical tests between individual utterance ratings might shed light on this question. Second, the silence duration used when delimiting IPUs in this study may have had an effect on the types of cues learned by our model. While there is no standard silence duration for IPU segmentation used in corpus studies, there is evidence silence threshold chosen can have an effect on the results obtained in corpus studies (Włodarczak and Wagner, 2013). For example, for turn-taking, we might find differing prosodic cues between IPUs which are followed by a long pause compared to IPUs followed by a very short pause. Third, we aimed to balanced the classes of turn-final and turn-medial utterances to ensure that there were sufficient turn-final samples in the training data. Thus by balancing the data, we are not evaluating the exact distribution of turn-position classes found in the data. For exploratory work this might not be a problem, and indeed oversampling from classes with fewer samples of turn-transition types has been used in previous corpus work e.g. Brusco et al. (2020). Finally, the data used in this work originated from online video-call interactions, which may affect the turn-taking cues used by speakers. An improvement of this method, therefore, would be to compare different corpora of conversational speech, including task-oriented dialogue. We would like to stress, that this method can be seen as a complementary approach to corpus studies.

For perceptual research, this method allows us to create stimuli from patterns learned directly from real conversational data. Moreover, this method allows us to control both the text being synthesised, as well as the speaker identity, which is desirable (Cutler and Pearson, 1985). In doing so, however, we relinquish control over the exact prosodic realisation. Future work could employ the approach introduced in Chapter 5, to create stimuli which differ on different prosodic dimensions, such as speaking rate, final $F_0$ contour and word duration, an approach similar to that of Lameris et al. (2024) for the role of creaky voice in turn-taking perception.

Finally, though we critiqued previous approaches to stimulus creation, which manipulate the prosody of natural recordings using splicing, or pitch manipulation, the data used in this work was not entirely optimal for producing very natural speech, as evidenced by the low MOS ratings. Nonetheless, the synthesis quality was sufficient for

listeners to perceive turn-taking cues learned from the CANDOR corpus, which shows the potential using found data in conjunction with speech synthesis models to explore linguistic questions. However, future work should validate the use of speech synthesis output in perceptual experiments more thoroughly. This could be done by replicating previous phonetic experiments using the output of current speech synthesis systems, or by comparing the naturalness of stimuli used in previous phonetic experiments made using traditional stimuli creation techniques such as splicing or $F_0$ manipulation with prosodic manipulation using speech synthesis methods. To create higher quality synthesis, cleaner conversational data could be also be used, for example data which has been elicited in corpus research in controlled recording environments.

## 7.7 Conclusion

In this work, we trained a TTS model with natural conversational data to model turn-taking cues. Overall, we found that our **TURN** model conditioned with the turn-final code was judged to sound more turn-final than the **TURN** model conditioned with the turn-medial code and more than the baseline, but results are speaker-specific. Interestingly, though our target speakers showed large imbalance in the number of training samples of turn-final utterances to turn-medial utterances, we were able to elicit turn-finality judgements in three of five speakers suggesting turn-finality cues can be learned from large amounts of data of many speakers. We have shown that TTS has potential to be used to analyse which cues listeners and speakers use in the context of turn-taking, though more analysis is needed.

# Part V

# Conclusion

# 8

# Discussion

## 8.1 Summary of Main Findings

The overarching goal of this thesis was to explore different aspects of context in conversational speech and investigate how these contextual features could be incorporated into speech synthesis models. However, conversational data is complex and there are a number of specific challenges that we faced when synthesising conversational speech, which were described in Chapter 1, but are summarised below:

| | |
|---|---|
| **Data** | We lack high-quality conversational speech datasets that are suitable for training speech synthesis models. |
| **Variation** | Conversational speech exhibits a tremendous amount of variation, both within and between speakers. Variation is also caused by the wide number of communicative contexts in which the speech occurs. |
| **Knowledge Gap** | Most research into conversational speech in speech science has been based on lab-elicited speech or corpus studies, and very few studies have examined the effect of context on the perception and production of speech in conversation |
| **Evaluation** | Speech synthesis evaluation has been developed for evaluating isolated utterances, we therefore lack standard paradigms for evaluating context-aware models and conversational speech. |

Due to the limitations described above, it became clear that to approach modelling conversational speech, we would have to address each of these limitations in different ways throughout the thesis. In each section, we present the main chapter(s) which addressed these issues.

### 8.1.1 Data

As we have seen, we lack high-quality conversational datasets for training speech synthesis models. Most of the recent work that has modelled contextualised conversational speech has used reenacted dialogue (Lee et al., 2023), or conversational datasets which are not publicly available (Guo et al., 2021; Yamazaki et al., 2020). In this thesis, we were interested in modelling naturally-occurring conversational speech, thus using a corpus such as DailyTalk (Lee et al., 2023) which is reenacted speech, or RyanSpeech Zandie et al. (2021), which comprises read *conversational-style* utterances, was not an option, nor were they available when this work commenced. In this thesis, we therefore used found data as a source of spontaneous naturally-occurring conversational speech.

#### Chapter 4

In Chapter 4, we addressed the data sparsity problem by using found podcast data to improve the prosody of a target speaker for whom we only had read speech utterances. Read speech differs from conversational speech in many respects, and one major aspect which differs across modalities is the distribution of certain linguistic constructions. For example, conversation contains dialogue acts which are not present, or are underrepresented in read speech corpora. We took *questions* as a case study. Questions are usually underrepresented in read speech corpora (Adigwe and Klabbers, 2022), but they are a fundamental group of dialogue acts in spoken dialogue systems. We showed that creating a training dataset by combining read speech and conversational speech led to significant increase in preference and MOS ratings for questions from our system, compared to the baseline system, which was only trained on read speech. Though we only focused on one aspect of conversation, this method could be applied to other linguistic structures which are present in conversation, but underrepresented in read speech data, such as backchannels, filled pauses or other speech acts. We also added statements from podcast data to our training set in the proposed model. However, we found a slight degradation in our system's performance compared to the baseline. We hypothesised that this could be due to the fact that read speech already contains a sufficient quantity of statements, thus adding more variable data from many speakers does not help. We conclude that this method may only benefit linguistic constructions which are underrepresented in the original read speech training corpus.

### 8.1.2 Variation

Conversational speech shows immense amounts of variation, and the sources of this variation are potentially infinite (Ogden, 2007) and are largely unknown. Variation can come from many different sources, for example, the speakers (e.g. their sociolect, dialect and idiosyncratic variation), the context that the speech occurs in (e.g. a formal meeting, a casual conversation or an interview), the previous conversational context and the intent of the speaker, just to name a few. Ideally, we want to account for as many sources of variation as possible – by doing this, we can begin to examine the contribution of these factors in conversation (Ogden, 2007). When synthesising conversational speech, accounting for variation is also crucial because it will improve prediction of acoustic features during training. In this thesis, we approached accounting for variation in conversational speech in two ways, which are described below.

### Chapter 6

In Chapter 6, we proposed a method to explore variation in a corpus of *found* conversational speech using a parametric representation of intonation that we introduced in Chapter 4. As in Chapter 3, we took a controlled approach by presenting a case study into the variation of the prosodic realisation of the Discourse Marker *well*. We chose to use DMs due to their multi-functionality in conversation, and in particular, due to their role in contextualising the target utterance in the wider conversational context. We presented a method to explore prosodic variation using clustering. We clustered intonational and durational features of the discourse marker, and found that clustering resulted in acoustically-related clusters and some pragmatically salient clusters (judged informally by the authors of the study). Though we only applied this method to a single word, the method can be used for any word or phrase in a conversational corpus. The benefit of using this method, is that it can allow us to extract prosodic patterns from naturally-occurring data and allows us to analyse interesting clusters, to explore TTS training data and to identify particular prosodic patterns that we may wish to model in conversational speech.

### Chapter 7

In Chapter 7, we took a different approach to accounting for variation in conversational data. Here, we investigated the potential of TTS as a complementary method to corpus research when we have access to a *known* source of variation. We used turn-taking as a case study for two reasons. First, turn-taking is one of the most fundamental aspects of conversation (Sacks et al., 1974) and the turn position of an utterance has been found to have associated prosodic correlates in previous corpus studies (e.g. Gravano and Hirschberg (2011); Brusco et al. (2017)) Second, when conversational speech is recorded in two channels, as it was in the CANDOR Corpus (Reece et al., 2023), the turn-position of an utterance can be automatically extracted.

We used speech synthesis as a tool to investigate the prosodic cues used in turn-taking. We trained a speech synthesis model which was additionally conditioned on the turn-position of each of our training utterances, which were extracted from the CANDOR Corpus. By conditioning a model on a turn-position embedding, we expected that this embedding would learn acoustic correlates of turn-position if they existed in the data. Thus, this allowed us to evaluate whether these feature exist *globally* in the data, similar to approaches in corpus studies. However, this method offered an additional benefit: by synthesising speech in each turn condition, we could evaluate whether such features were perceptible by listeners. In our results, we found that synthesising the turn-conditioned model with the turn-final tag led to a significant increase in turn-finality judgements over the baseline condition and over the speech synthesised with the turn-medial tag. This result was found for three of the five speakers used in evaluation. This suggests that the model had learned turn-taking patterns from the training data, but that this effect potentially interacts with the speaker's prosodic characteristics.

One limitation of this study is that not all model speakers showed an increase in performance in turn-finality judgements. This might be due to the fact that some speakers realised their utterances with *uptalk*. As we have alluded to previously, this shows that sociophonetic and dialectal information will also impact the exploration of prosodic cues in conversational speech. It is therefore vital that sociophonetic variation is accounted for,

as different dialects show differences in their use of prosodic cues (e.g. Grabe et al. (2005)). Thus, future work should include more identification of speaker accent, gender and sociolect, which are all likely to affect the results in this work.

### 8.1.3 Knowledge Gap

In both speech science and speech synthesis, work has largely concentrated on studying the acoustic features of isolated utterances. We still lack fundamental understanding on how speech is used in every day interactions and how the interactive and pragmatic context affects both our perception of speech and our production of subsequent turns. One of the reasons why studying contextualised conversational speech remains a challenge, in particular for speech perception research, is that it is extremely difficult to create spontaneous, but controlled, conversational stimuli. Due to this, previous work has used prosodically-manipulated conversational speech, or reenacted conversational speech using heavily-controlled textual material, however this might not be ecologically valid. In this thesis, we developed methods to create stimuli from patterns found in real conversational data, which were subsequently synthesised with speech synthesis. These methods are described below.

#### Chapter 5

In this chapter, we presented a method for controlling an end-to-end speech synthesis model with linguistically-relevant features representing intonation. While many modern end-to-end models employ various methods of controllability, many of these methods are not linguistically-motivated or interpretable. For speech synthesis models to be of use in speech science, we believe that the representations used to control the models should link back to representations used in prosodic theory. For this reason, we chose to parameterise $F_0$ using Legendre polynomials, which have been used in many previous linguistic studies. We demonstrated that using a data-driven hierarchical specification of the $F_0$ contour on the phrase-level using slope and on prominent words using Legendre polynomial coefficients led to synthetic speech that was more similar to a reference recording than the baseline and a categorically-conditioned model. Further, in Chapter 6, we demonstrated how this model can be used to create stimuli from prosodic patterns found in conversational data. Using prosodic patterns from conversational data allows us to create stimuli with acoustic features that are attested in natural speech. This is in contrast to post-processing methods, such as TD-PSOLA which are often used to create stimuli by imposing contours based on top-down knowledge of prosody or are constructed by the experimenter based on features attested in the literature.

### 8.1.4 Evaluation

Speech synthesis was developed for isolated utterances, and therefore speech synthesis evaluation paradigms have not been developed for contextualised utterances. Furthermore, common metrics which are used in traditional evaluation paradigms include *naturalness* and *intelligibility*. When evaluating conversational speech, however, we are more interested in prosodic variation and in-context evaluation. In this thesis, we therefore investigated whether the MOS paradigm is suitable for rating speech in context.

**Chapter 3**

Clark et al. (2019) found that synthetic speech was rated more highly in context than in isolation. We investigated why this effect occurs by investigating the effect of task instructions, the textual context-dependence between utterances, and the prosodic felicity of utterances. We found that the task instructions had a significant effect on the MOS ratings. When asked to rate *naturalness* in isolation or in context, no significant difference between ratings was found, but asking listeners to rate *appropriateness* of utterances presented in context resulted in a rating higher than the naturalness score for isolated utterances. This replicated the findings in Clark et al. (2019). Naturalness and appropriateness are not interchangeable concepts. This highlights the importance of reporting experiment instructions used in evaluation studies.

In this study, however, we were specifically interested in whether the MOS paradigm is appropriate for rating prosodically varied speech from the same system. We found that when utterances were presented in isolation, participants exhibited a greater preference for canonically realised renditions, a preference that was maintained for utterances presented in context, but to a lesser extent. The fact that the ratings between the canonically-realised and non-canonically-realised utterance became closer in context, suggests that MOS may not be sensitive to subtle prosodic differences. More recent research by Camp et al. (2023), has suggested that side-by-side comparison is more suitable for system comparison. The results of this work motivated our controlled and targeted approach to evaluation, which was used in the case studies of Chapter 6 and Chapter 7. In Chapter 6, we evaluated a single contextual effect, namely the speaker stance, while in Chapter 7 we evaluated turn-finality.

## 8.2 Reflections and Future Work

One of the main motivations for the interdisciplinary approach to studying conversational speech taken in this thesis was that from both a speech science and speech synthesis standpoint, conversational speech is a challenging domain, and we still know relatively little about it compared to *lab speech*. In speech science, we have historically focused our analyses on isolated utterances of read or lab speech (Wagner et al., 2015b), and in speech synthesis research we have also trained our models on such utterances, but why speakers realise each utterance in a conversation the way they do under certain pragmatic conditions still remains elusive. In the following final two sections of this thesis, we will reflect on how the methods presented in this thesis can be used in both speech science and speech technology, in particular focusing on further directions of research that should be pursued.

### 8.2.1 Speech Science

For speech science, to study conversation we need to move from exclusively studying laboratory speech to utilising more diverse datasets of real human interaction. However, as noted by Wagner et al. (2015b), in doing so, we are relinquishing control over the form of the speech and the content of what is being said, which brings its own challenges, such as not being able to isolate enough instances of certain spoken phenomena (Wagner et al., 2015b), and consequently making it more difficult (though not impossible see Wallbridge et al. (2021)) to set up perception experiments using speech from such datasets because speech that has been elicited in a more naturalistic setting will likely not have enough

instances of textually identical utterances spoken in different ways in particular contexts by the same speaker. In this thesis, we therefore present speech synthesis as a tool which can aid speech scientists in working with natural conversational data, presenting work to allow for the exploration of variation across speakers, but also allowing us to regain a level of control in stimuli creation by providing a method to create experimental stimuli from patterns found in *real* data using speech synthesis.

However, for use in speech science research, there are a number of future directions that will need to be taken. First, more work needs to be done to validate the use of speech synthesis output in experimental phonetics before it can be heavily adopted in the field. For example, the method of speech synthesis controllability presented in Chapter 4, while allowing us to control the speech synthesis output using more linguistically-relevant units, does not provide perfect control of the $F_0$ of an utterance. As we saw, the same will of course be true for real speakers, who will not be able to mimic the exact characteristics of a reference audio, therefore we need to validate whether the speech output of the controlled model falls into the correct range that would be expected in a mimicry task, and whether it indeed produces the intended contour (something RMSE and correlation cannot *directly* tell us).

Similarly, we motivated the use of speech synthesis in speech science research by the fact that the quality of speech synthesis is reaching high levels of naturalness. However, this might not be the case when training on more diverse conversational data and found data which has been recorded in many different recording environments (see Chapter 7 in particular). We therefore need to evaluate the external validity of using synthetic speech in perceptual research, especially since previous research has found differences in the cognitive load of listeners when listening to synthetic speech compared to natural speech (Govender et al., 2019). To do this, a further line of inquiry would be to replicate well-known perception experiments, such as the perceptual learning paradigm (Norris et al., 2003), using speech synthesis output and subsequently comparing this to more common methods of creating stimuli, such as recording natural stimuli, but also traditional forms of stimuli manipulation, such as splicing and TD-PSOLA, which are also often used in speech science research.

A further limitation of the method presented in this thesis is that we have only taken English as our target language. For wide-spread adoption in speech science, it would be helpful to apply these methods to multiple languages. Therefore, to assess the use of the methods presented in this thesis, future work should apply these methods to a wider range of languages. Similarly, for under-resourced languages, there may not be enough data or linguistic resources available for training the models that were used in this thesis. In particular, for conversational speech, we used various models to filter the data, such as ASR models for generating improved transcripts and other models for detecting laughter or overlapping speech. These auxiliary models which were used to create the training data for our experiments may not have been trained on a wide variety of languages and this might therefore limit their performance for other languages. This of course places a limit on how much the methods in this thesis can be applied to a wider set of languages.

Finally, in this thesis we also presented methods for exploring prosodic variation in real conversational corpora using clustering. Clustering is being applied in more and more prosodic research (see Calhoun and Schweitzer (2012); Zellers and Ogden (2014); Cole et al. (2024)), and similar to the work in this thesis, it has also been applied to found data, for example in Aviad et al. (2024). Such exploratory analysis is an important step in discovering the patterns of actual language use in conversation, which may differ from

what is assumed in the literature (Cangemi et al., 2023). For the work presented in this thesis, and especially in Chapter 6 in which we explored the prosody of the discourse marker *well*, future work should validate the perceptual similarity between items in each cluster. Validating the content of the clusters, and performing perception tests directly on the real speech found in the clustering exploration could be a fruitful line of research, especially when the appropriate number of clusters is unknown. Further, applying this method to a subset of labelled data from a linguistically-curated corpus would be an important step towards validating the method.

### 8.2.2 Speech synthesis

For speech synthesis research, we largely suffer from the same issues that we face in speech science when working with conversational speech, namely we are confronted with the overwhelming variation in naturally-occurring conversational data and we currently have no standardised methods to evaluate conversational speech or contextualised speech. In the same way that speech synthesis can be used as a tool in speech science, speech scientists can help to explore these issues and develop new methods to evaluate speech synthesis models, as well as evaluate what aspects of conversational speech should be included in a speech synthesis model.

First, though there has been increased focus on synthesising conversational speech and evidence that more spontaneous-sounding speech is rated as more suitable for situations in which spontaneous speech is more likely to occur, such as casual conversations (Székely et al., 2019b) and more natural when no definition of naturalness is given (Dall et al., 2014b), future research should begin to explore what styles of speech are more appropriate in different applications from the perspective of the end user. In particular, future work should focus on how realistic the output of a speech synthesis model should sound and to what extent human-like conversational behaviour related to the spontaneous nature of speech production in conversation should be replicated. Examples of such human behaviours include false starts, laughter, breathing, and emotional cues, but also prosodic-pragmatic cues, such as using prosody to signal stance or attitude. Therefore more research, similar to the recent work of Ross et al. (2024) who explored the uncanny valley in synthetic speech, is needed to explore how natural or human-like the speech has to be to be rated positively and whether human-likeness should be the gold standard. Moreover, future work should also evaluate the ethical considerations of creating human-like voices which are indistinguishable from human speech.

Second, in this thesis, we focused on very specific and targetted evaluation of prosody and conversational speech synthesis in the case studies of Chapter 6, which explored the perceived agreement expressed in the speech, and Chapter 7 which explored the use of prosody in conveying end of turn. However, for most speech synthesis applications constructing such detailed and focused evaluation paradigms might not be cost- or time-effective. Ideally, future work should aim to create conversational test sets created to capture a wide range of conversational behaviour that a model should be able to produce, for example asking questions with the correct prosody, changing prosodic prominence based on context or signalling turn-taking cues, just to name a few. Furthermore, the evaluation in this thesis was still based on non-interactive scenarios, removed from the communicative context that these utterances should have been embedded in. For applications where conversational speech is desirable, we need to evaluate in that context as

this has been shown to affect ratings of speech (Lameris et al., 2023). For example, for our turn-taking model from Chapter 7, we could ask whether employing a turn-aware model aids the turn management in a conversation between a human and a spoken dialogue system. Future work should therefore embed the conversational synthetic speech in a more realistic use case during evaluation (Wagner et al., 2019).

Finally, one of the main limitations in this work, regarding speech synthesis modelling, is that we have focused nearly exclusively on the FastPitch (Łańcucki, 2021) architecture. During the time that this thesis was carried out, speech synthesis models and methods have improved. For example, as we saw, the use of self-supervised units instead of canonical transcriptions (Wells et al., 2023) might be able to mitigate the issues with mismatched phonetic realisation between read and spontaneous speech, allowing us to better model phonetic reduction. Further, other speech synthesis architectures have become available, such as VITS (Kim et al., 2021), X-TTS (Casanova et al., 2024) and in particular for synthesising spontaneous speech MQTTS (Chen et al., 2023). It is therefore important to evaluate whether the methods described in this thesis are applicable to newer architectures and for the purposes of speech science research, whether newer architectures can increase the performance and external validity of using speech synthesis in the phonetic sciences.

# A

# Appendix

## A.1 Chapter 3: Textual Material for Listening Tests

Table A.1: *Textual Material used in Listening Tests for Chapter 3 (Contexts)*

| **Stimulus** | **Context** |
|---|---|
| ex_context_01 | Gorillas may be a lot hairier than we are, but they are our third closest relatives. |
| ex_context_02 | Santiago is the capital of Chile and is one of the largest cities in South America. |
| ex_context_03 | Contrary to popular belief, woodlice are not insects but are in fact crustaceans. |
| ex_context_04 | Ireland is an island country in the North Atlantic with a population of nearly five million people. |
| ex_context_05 | Rwanda, one of the smallest African nations, is located a few degrees south of the equator. |
| ex_context_06 | Oranges are a delicious and incredibly popular fruit eaten all across the world. |
| ex_context_07 | Dutch is one of the closest relatives to both English and German. |
| ex_context_08 | Pavlova is a popular dessert consisting of meringue, cream and different fruits and toppings. |
| ex_context_09 | Monet was one of the most famous artists of the impressionism movement. |
| ex_context_10 | Common toads might not be the prettiest animals, but they can live up to a staggering fifty years in captivity. |
| ex_context_11 | You might think of witches as a product of fiction, commonly found in children's books. |
| ex_context_12 | Germany is famous for its production of many different items, but cheese might not be the first you think of. |
| ex_context_13 | Denali, the third highest peak in the world, is found in the Denali National Park in Alaska. |
| ex_context_14 | No language is more difficult to learn than the other, but there are reasons why some can be more difficult. |
| ex_context_15 | Go is one of the oldest boardgames which is still played today. |

| Stimulus | Context |
|---|---|
| ex_context_16 | Vienna has one of the best networks of public transport in Europe, and one of the cheapest. |
| ex_context_17 | Macbeth, one of Shakespeare's tragedies was first performed in sixteen o six. |
| ex_context_18 | Lying along the banks of the River Nile in North Africa, lie the ruins of the Egyptian's ancient civilisation. |
| ex_context_19 | Nightshades are a family of plants, some of which are highly toxic, some of which are edible and extremely tasty. |
| ex_context_20 | Dosas are a thin pancake dish eaten with sambar or chutney in South India. |
| ex_context_21 | Kenya is an east African country with a population of nearly forty eight million people. |
| ex_context_22 | Eric Arthur Blair, probably best known as George Orwell, was an English novelist and essayist. |
| ex_context_23 | Normal People, the two thousand and eighteen novel by writer Sally Rooney has recently been made into a series. |
| ex_context_24 | Alexandria Ocasio Cortez, known as AOC, is a member of the house of representatives representing New York. |
| ex_context_25 | Mastermind is a popular general knowledge quiz show which airs on the BBC. |
| ex_context_26 | We often think of bears as being aggressive animals, which can be quite dangerous to humans. |
| ex_context_27 | Urban dictionary is a website providing a dictionary service for slang terms in the English language. |
| ex_context_28 | Chlorophyll from the Greek for green leaf, is a green pigment found in plants. |
| ex_context_29 | Olives are an ancient fruit, and some of its trees are more than three thousand years old. |
| ex_context_30 | The sport of rugby union was allegedly the invention of a boy named Webb Ellis. |
| ex_context_31 | Storms have been named in the US since the seventeen hundreds, for the UK it's a relatively new thing. |
| ex_context_32 | Many religions have special rules for what a religious member may or may not eat, one such religion is Jainism. |
| ex_context_33 | Relish is a type of condiment which normally consists of vegetables or fruits in a sauce. |
| ex_context_34 | Berlin has been the capital city of Germany since reunification in nineteen ninety. |
| ex_context_35 | A favourite food of many around the world, chocolate creates mildly stimulating effects. |
| ex_context_36 | Depression is a common psychological disorder and is the leading cause of disability around the world. |
| ex_context_37 | Malted barley is the most common ingredient in beer, a fermented alcoholic beverage. |
| ex_context_38 | The word cake originated from the Vikings who spoke Old Norse. |
| ex_context_39 | Hurling is an ancient team sport played in Ireland and is related to Scottish shinty. |
| ex_context_40 | The Dutch might lay claim to being the largest cultivators of the tulip, but they were not the first. |
| ex_context_41 | Loganberries are a hybrid fruit of the blackberry and raspberry and are red in colour. |

| Stimulus | Context |
|---|---|
| ex_context_42 | The kindle is an e-book reader produced by Amazon, and features a large library available for download. |
| ex_context_43 | Bicycles may not be known for changing history, but have done more than you would think. |
| ex_context_44 | Chilis are a fruit of the plant genus capsicum, and their taste ranges from mild to spicy. |
| ex_context_45 | Aquafaba, deriving from the latin words for water and beans, is the leftover water from cooking beans. |
| ex_context_46 | Turtles are very popular pets, but their living arrangements may shock you. |
| ex_context_47 | Whisky has long been a staple of Scottish produce with distilleries in all corners of the country. |
| ex_context_48 | Tempeh is a soy based product, and is made by fermenting soybeans. |
| ex_context_49 | Pidgin languages form in language contact situations where languages are mixed in order to communicate. |
| ex_context_50 | A key late-stage symptom of the viral disease rabies is hydrophobia. |
| ex_context_51 | Rosalind Franklin was a British chemist whose work was pivotal in the discovery of DNA structures. |
| ex_context_52 | Insulin is a hormone needed by humans, and is absent or reduced when a person has diabetes. |
| ex_context_53 | Despite their name, Danish pastries did not in fact originate in Denmark at all. |
| ex_context_54 | Since the twentieth century, there has been increased interest in women's football, both participation and spectation. |
| ex_context_55 | Screen printing is a method used to transfer ink onto a material. |

Table A.2: *Textual Material used in Listening Tests for Chapter 3 (Targets context-dependent)*

| Stimulus | Target |
|---|---|
| ex_target_01 | We share more than ninety five percent of our DNA with them. |
| ex_target_02 | It was founded in fifteen forty one and its population is just over six million. |
| ex_target_03 | Unlike related species, crabs and lobsters, they are said to have an unpleasant taste. |
| ex_target_04 | Its two official languages are Irish, a Celtic language, and English. |
| ex_target_05 | It is one of three countries to have a female majority in government. |
| ex_target_06 | Roughly seventy million tonnes of them are grown around the world each year. |
| ex_target_07 | It is not only the official language of the Netherlands, but is also spoken in parts of Belgium. |
| ex_target_08 | For years both New Zealand and Australia have fought to be seen as the dessert's creator. |
| ex_target_09 | The term impressionism actually originated from the title of his work, Impression Sunrise. |
| ex_target_10 | In the wild, they live significantly shorter than fifty years, at around ten to twelve years. |
| ex_target_11 | Interestingly, they are found in various cultures around the world, for example in the Philippines. |
| ex_target_12 | It actually produces zero point one million tonnes more cheese than France. |

| Stimulus | Target |
|---|---|
| ex_target_13 | Some might know it as Mount McKinley, but it was renamed in twenty fifteen following a long dispute. |
| ex_target_14 | One factor, which can predict ease of learning a new one, is linguistic similarity to your native language. |
| ex_target_15 | It originated in China and is more than two thousand five hundred years old. |
| ex_target_16 | Its yearly transport ticket works out to be only one euro per day. |
| ex_target_17 | It starts with the spooky scene of the three witches who tell him he will be king of Scotland. |
| ex_target_18 | While famous for the Pyramids of Giza, they were also known for systems of mathematics and medicine. |
| ex_target_19 | You have probably eaten some of them before, including tomatoes, aubergines and potatoes. |
| ex_target_20 | They are made of fermented batter made from ground lentils and rice. |
| ex_target_21 | It gained independence from the United Kingdom in nineteen sixty three. |
| ex_target_22 | Not only was he responsible for the phrase Big Brother is watching you, he actually coined the term Cold War. |
| ex_target_23 | The series received critical acclaim and had more than sixteen million views in less than a month. |
| ex_target_24 | She is the youngest woman to serve in the united states congress since its founding in seventeen forty eight. |
| ex_target_25 | It was inspired by its founders experience of being interrogated by the Gestapo in the second world war. |
| ex_target_26 | They are not all aggressive, polar bears are actually more inclined to flight rather than fight. |
| ex_target_27 | While this dictionary doesn't contain entries typically found in a normal dictionary, it's proven useful in documenting changing language use. |
| ex_target_28 | It absorbs blue and red wavelengths, but not green light which causes plants to appear green in colour. |
| ex_target_29 | Fossil evidence suggests that they might have been around for over twenty million years. |
| ex_target_30 | International teams of this sport play in the world championship competing for a trophy in Webb Ellis' name. |
| ex_target_31 | The first one to receive a name in the UK was storm Abigail in two thousand and fifteen. |
| ex_target_32 | Besides being vegetarian, members are not allowed to eat root vegetables such as onions and garlic. |
| ex_target_33 | In the United States the most common one is made with diced pickles. |
| ex_target_34 | It is now the most populous city in the European Union since the UK departure from the union. |
| ex_target_35 | A normal sized serving of it contains about as much caffeine as a decaf coffee, but darker variants contain more. |
| ex_target_36 | It can lead to higher prevalence of other diseases such as dementia, as well as increased physical ailments. |
| ex_target_37 | Compared to other cereals it contains large amounts of gluten, so it is not always the best choice. |

| Stimulus | Target |
|---|---|
| ex_target_38 | It is not the only word to come from Old Norse, other examples include Thursday, husband and blunder. |
| ex_target_39 | It is one of the fastest field sports played in the world and is highly complicated. |
| ex_target_40 | The plant's earliest known cultivation was in Persia, most probably in the tenth century. |
| ex_target_41 | They were used by the British navy to prevent scurvy due to their high Vitamin C content. |
| ex_target_42 | Many users appreciate the compact nature of the device, but occasionally miss the feel of a real book. |
| ex_target_43 | They interestingly helped female emancipation by increasing women's mobility in the eighteen hundreds. |
| ex_target_44 | The spiciest one ever recorded is the Carolina Reaper according to the Guinness book of records. |
| ex_target_45 | It has risen in popularity as a substitute for eggs in foods like vegan mayo and meringue. |
| ex_target_46 | From mid-November until the end of winter, they are often kept in the fridge to hibernate. |
| ex_target_47 | There exists more barrels of it in Scotland, than there are people residing in the country. |
| ex_target_48 | It has gained popularity around the world but is a staple on the island of Java. |
| ex_target_49 | When these languages are acquired by children as their first language, they become known as creole languages. |
| ex_target_50 | This literally means fear of water and manifests as panic when confronted with water and difficulty swallowing. |
| ex_target_51 | Rosalind Franklin didn't receive the Nobel prize for the discovery of its structures, but her work made it possible. |
| ex_target_52 | Dorothy Hodgkin is responsible for discovering its structure enabling its mass production for diabetics. |
| ex_target_53 | The pastries originated in Vienna and the idea was passed on to the Danes by Austrian bakers. |
| ex_target_54 | Unfortunately, as in many other sports, pay and opportunities are much lower in this sport. |
| ex_target_55 | This method was popularised in the sixties during the pop art movement. |

Table A.3: *Textual Material used in Listening Tests for Chapter 3 (Targets context-independent)*

| Stimulus | Target - stand alone version |
|---|---|
| ex_target_sa_01 | We share more than ninety five percent of our DNA with Gorillas. |
| ex_target_sa_02 | Santiago was founded in fifteen forty one and its population is just over six million. |
| ex_target_sa_03 | Unlike related species, crabs and lobsters, woodlice are said to have an unpleasant taste. |
| ex_target_sa_04 | Ireland's two official languages are Irish, a Celtic language, and English. |
| ex_target_sa_05 | Rwanda is one of three countries to have a female majority in government. |

| Stimulus | Target - stand alone version |
|---|---|
| ex_target_sa_06 | Roughly seventy million tonnes of oranges are grown around the world each year. |
| ex_target_sa_07 | Dutch is not only the official language of the Netherlands, but is also spoken in parts of Belgium. |
| ex_target_sa_08 | For years both New Zealand and Australia have fought to be seen as the Pavlova's creator. |
| ex_target_sa_09 | The term impressionism actually originated from the title of Monet's work, Impression Sunrise. |
| ex_target_sa_10 | In the wild, toads live significantly shorter than fifty years, at around ten to twelve years. |
| ex_target_sa_11 | Interestingly, witches are found in various cultures around the world, for example in the Philippines. |
| ex_target_sa_12 | Germany actually produces zero point one million tonnes more cheese than France. |
| ex_target_sa_13 | Some might know Denali as Mount McKinley, but it was renamed in twenty fifteen following a long dispute. |
| ex_target_sa_14 | One factor, which can predict ease of learning a new language, is linguistic similarity to your native language. |
| ex_target_sa_15 | Go originated in China and is more than two thousand five hundred years old. |
| ex_target_sa_16 | Vienna's yearly transport ticket works out to be only one euro per day. |
| ex_target_sa_17 | Macbeth starts with the spooky scene of the three witches who tell him he will be king of Scotland. |
| ex_target_sa_18 | While famous for the Pyramids of Giza, Egyptians were also known for systems of mathematics and medicine. |
| ex_target_sa_19 | You have probably eaten some Nightshade plants before including tomatoes, aubergines and potatoes. |
| ex_target_sa_20 | Dosas are made of fermented batter made from ground lentils and rice. |
| ex_target_sa_21 | Kenya gained independence from the United Kingdom in nineteen sixty three. |
| ex_target_sa_22 | Not only was Orwell responsible for the phrase Big Brother is watching you, he actually coined the term Cold War. |
| ex_target_sa_23 | Normal People received critical acclaim and had more than sixteen million views in less than a month. |
| ex_target_sa_24 | AOC is the youngest woman to serve in the united states congress since its founding in seventeen forty eight. |
| ex_target_sa_25 | Mastermind was inspired by its founders experience of being interrogated by the Gestapo in the second world war. |
| ex_target_sa_26 | Bears are not all aggressive, polar bears are actually more inclined to flight rather than fight. |
| ex_target_sa_27 | While Urban Dictionary doesn't contain entries typically found in a normal dictionary, it's proven useful in documenting changing language use. |
| ex_target_sa_28 | Chlorophyll absorbs blue and red wavelengths, but not green light which causes plants to appear green in colour. |
| ex_target_sa_29 | Fossil evidence suggests that olives might have been around for over twenty million years. |
| ex_target_sa_30 | International teams of rugby union play in the world championship competing for a trophy in Webb Ellis' name. |

| Stimulus | Target - stand alone version |
|---|---|
| ex_target_sa_31 | The first storm to receive a name in the UK was storm Abigail in two thousand and fifteen. |
| ex_target_sa_32 | Besides being vegetarian, Jains are not allowed to eat root vegetables such as onions and garlic. |
| ex_target_sa_33 | In the United States the most common relish is made with diced pickles. |
| ex_target_sa_34 | Berlin is now the most populous city in the European Union since the UK departure from the union. |
| ex_target_sa_35 | A normal sized serving of chocolate contains about as much caffeine as a decaf coffee, but darker variants contain more. |
| ex_target_sa_36 | Depression can lead to higher prevalence of other diseases such as dementia, as well as increased physical ailments. |
| ex_target_sa_37 | Compared to other cereals barley contains large amounts of gluten, so it is not always the best choice. |
| ex_target_sa_38 | Cake is not the only word to come from Old Norse, other examples include Thursday, husband and blunder. |
| ex_target_sa_39 | Hurling is one of the fastest field sports played in the world and is highly complicated. |
| ex_target_sa_40 | The tulip's earliest known cultivation was in Persia, most probably in the tenth century. |
| ex_target_sa_41 | Loganberries were used by the British navy to prevent scurvy due to their high Vitamin C content. |
| ex_target_sa_42 | Many users appreciate the compact nature of the kindle, but occasionally miss the feel of a real book. |
| ex_target_sa_43 | Bicycles interestingly helped female emancipation by increasing women's mobility in the eighteen hundreds. |
| ex_target_sa_44 | The spiciest chili ever recorded is the Carolina Reaper according to the Guinness book of records. |
| ex_target_sa_45 | Aquafaba has risen in popularity as a substitute for eggs in foods like vegan mayo and meringue. |
| ex_target_sa_46 | From mid-November until the end of winter, Turtles are often kept in the fridge to hibernate. |
| ex_target_sa_47 | There exists more barrels of Whisky in Scotland, than there are people residing in the country. |
| ex_target_sa_48 | Tempeh has gained popularity around the world but is a staple on the island of Java. |
| ex_target_sa_49 | When pidgin languages are acquired by children as their first language, they become known as creole languages. |
| ex_target_sa_50 | Hydrophobia literally means fear of water and manifests as panic when confronted with water and difficulty swallowing. |
| ex_target_sa_51 | Rosalind Franklin didn't receive the Nobel prize for the discovery of DNA structures, but her work made it possible. |
| ex_target_sa_52 | Dorothy Hodgkin is responsible for discovering insulin's structure enabling its mass production for diabetics. |
| ex_target_sa_53 | Danish pastries originated in Vienna and the idea was passed on to the Danes by Austrian bakers. |

| **Stimulus** | **Target - stand alone version** |
|---|---|
| ex_target_sa_54 | Unfortunately, as in many other sports, pay and opportunities are much lower in women's football. |
| ex_target_sa_55 | Screen printing was popularised in the sixties during the pop art movement. |

## A.2 Chapter 4: Textual Material for Listening Tests

Table A.4: *Textual Material used in Listening Tests for Chapter 4 (Answers)*

| Stimulus | Text |
|---|---|
| answer_81 | No one has ever. |
| answer_85 | I think they'll stay for another year. |
| answer_75 | I think we've got a lot better at this but there's still some way to go. |
| answer_5 | Yeah, but that's what I mean. |
| answer_76 | I am going to be talking about Akron. |
| answer_83 | Oh man the demo team five times. |
| answer_29 | So now you put me on the spot. |
| answer_40 | I love this week. |
| answer_68 | No, I'd be happy to help regardless of the result. |
| answer_63 | That's a really good question. |
| answer_52 | I'm going to try to help you guys out. |
| answer_62 | You should watch this film Chris and I'm calling you out Chris. |
| answer_96 | I'm great and Martin. |
| answer_51 | I do I do. |
| answer_41 | Oh, of course, I think you know regardless of what we've seen over the last two weeks. |
| answer_88 | Ah good question. |
| answer_80 | So it's more of an evening thing. |
| answer_7 | Not particularly. |
| answer_46 | I'm still trying to figure that out. |
| answer_37 | I'm not Meg. |
| answer_17 | You know, I'm actually quite hidden on social media. |
| answer_87 | Just work everyday train every day. |
| answer_27 | You know that you guys are gonna ask me. |
| answer_48 | Yeah, so you can follow me on Twitter. |
| answer_94 | Exactly exactly. |
| answer_25 | Well, it depends how good you are. |
| answer_86 | I mean, I'm from Yellow Springs, Ohio. |
| answer_84 | I love to challenge the status quo. |
| answer_79 | Yeah, no definitely. |
| answer_49 | Yeah, exactly and long. |
| answer_65 | Yeah, man, it was frustrating. |
| answer_82 | So there's that. |
| answer_92 | You know, I've gotten really used to it. |
| answer_20 | That's a good question. |
| answer_32 | I know no details of it. |
| answer_47 | I think it's a good answer. |
| answer_45 | Oh, I think so. |
| answer_61 | Yeah, Weasley's kitchen. |
| answer_33 | Sure, not just photos right now. |
| answer_9 | Yeah, like when I was, you know a freshman sophomore. |
| answer_21 | Hope is not a strategy. |

| | |
|---|---|
| answer_44 | Well, she was there in in the crowd. |
| answer_42 | Come on, come on genius. |
| answer_66 | I don't know. |
| answer_89 | Yeah, like being a liberal person but specifically like liberal fashion-wise. |
| answer_78 | I would love to go to the Caribbean. |
| answer_26 | It's really good. |
| answer_97 | No, it's tiger. |
| answer_28 | Well Luke cohabitating is basically what we've been doing for the past two years, which is living together. |
| answer_39 | No still keep them in check. |

Table A.5: *Textual Material used in Listening Tests for Chapter 4 (Questions)*

| **Stimulus** | **Text** |
|---|---|
| question_31 | And so what are you doing exactly like right now? |
| question_80 | What made him actually swing that bat right off the bat on him? |
| question_63 | How is your relationship with her? |
| question_36 | Do you write a lot? |
| question_3 | What do you think? |
| question_66 | As and why did I get that done? |
| question_92 | How is your summer break going at the moment? |
| question_75 | How close was he to becoming the third brother of Destruction? |
| question_59 | How do I get better on my craft? |
| question_69 | Is that the plans? |
| question_67 | You remember your first time? |
| question_64 | Did you tell her happy birthday? |
| question_43 | Are you sure you didn't change your name? |
| question_83 | How do you think the team is going to fare going into the sea games? |
| question_72 | Did you plan you know to be an activist that day or did it just happen? |
| question_6 | Why enjoy? |
| question_65 | Did you know that it was coming? |
| question_44 | So you really really want to avoid merging at rainless company? |
| question_14 | How could they have survived in such a difficult landscape? |
| question_82 | Did you ever watch Jersey Shore? |
| question_28 | But yeah, I say, where's my wife? |
| question_52 | Does that take the cake on your dumbness? |
| question_88 | How did that come about? |
| question_26 | Do we have sponsors? |
| question_81 | How hard would it be to choose which guards are there? |
| question_76 | Do we have to like kill the next? |
| question_71 | Is the stupidest game the room? |
| question_56 | Like what were they doing? |
| question_74 | What did you think about the girls being at the mercy of Boomer? |
| question_58 | Rory do you wear jeans? |
| question_87 | What do you look for in a challah? |
| question_19 | What do you do to check out? |

| | |
|---|---|
| question_34 | Is it when it came back out in the remake? |
| question_55 | What made you have that decision? |
| question_51 | So do you think you're one of these people that would hold a grudge? |
| question_61 | So what's going on in your world? |
| question_35 | They can find you and email you with what do they do? |
| question_7 | So would you say that life is fair? |
| question_22 | How do you feel about him turning seven? |
| question_47 | You know what you need to do? |
| question_45 | So who are you at the end of the day? |
| question_86 | So what do you think Damian Lillard can possibly do for an encore? |
| question_73 | Oh, what's coming out? |
| question_27 | Can you name any of the stores? |
| question_93 | What sparked that? |
| question_29 | In-between stage? |
| question_39 | What was the best day of your life? |
| question_9 | Do you have anything else to add offensively? |
| question_70 | Why not do it, right? |
| question_10 | Wonder how about you any closing thoughts? |

## A.3 Chapter 4: Supplementary Figures

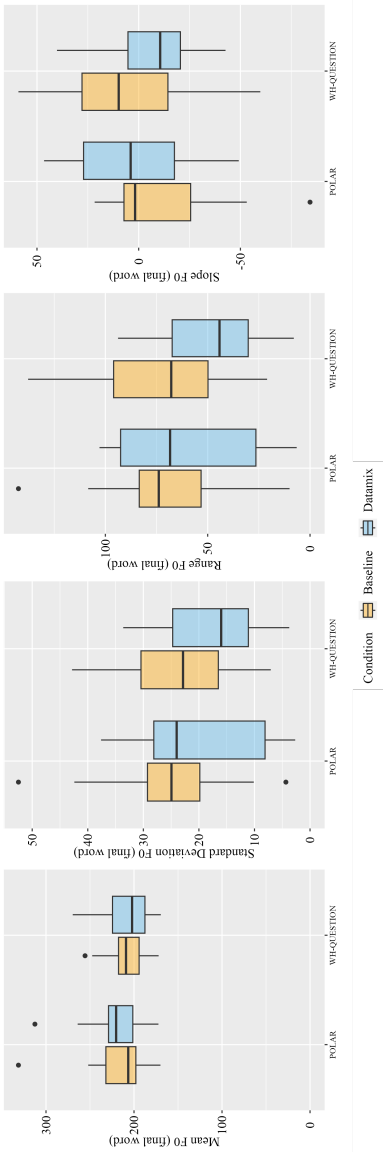


Figure A.1: *Boxplots of prosodic features of final word for questions*

## A.4 Chapter 5: Supplementary Figures

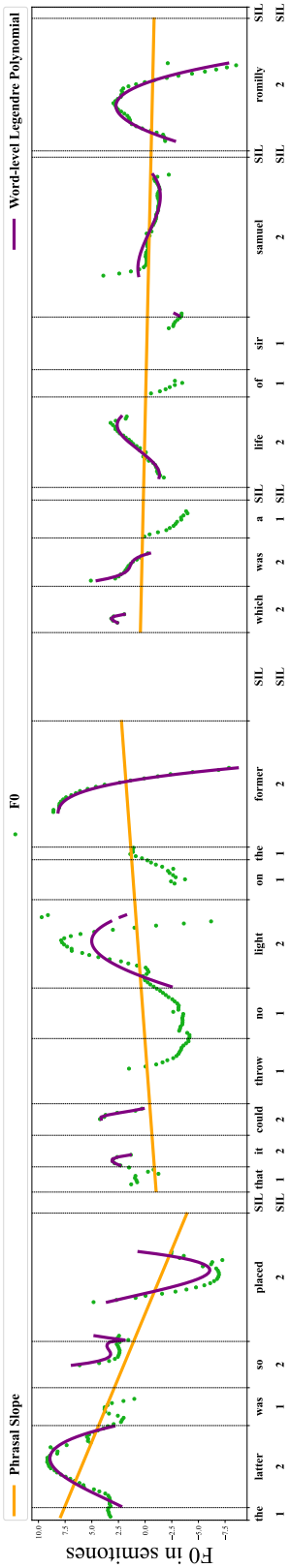


Figure A.2: $F_0$ *with fitted slope and Legendre Polynomials LJ013-0179. The x-axis shows both word alignments and prominence category.*

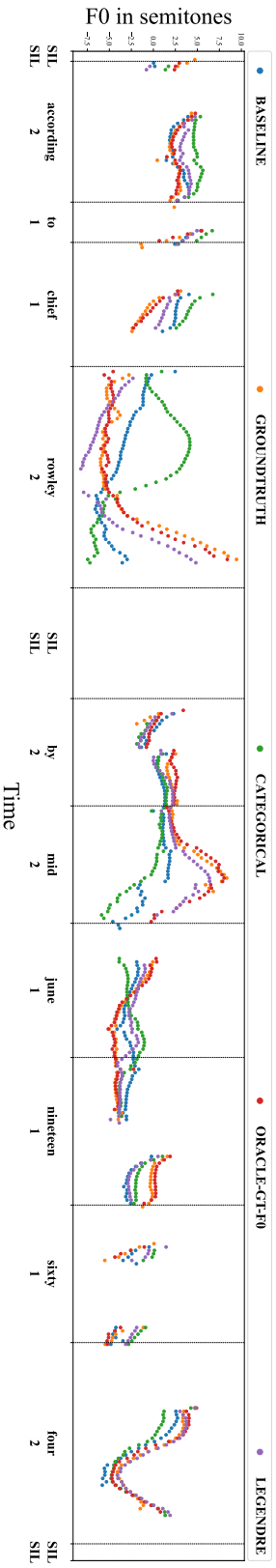


Figure A.3: *$F_0$ contours of all models and ground truth for utterance LJ050-0068. $x$-axis shows both word alignments and which words received a prominence (and therefore a set of Legendre polynomial coefficients).*

## A.5 Chapter 7: Supplementary Figures

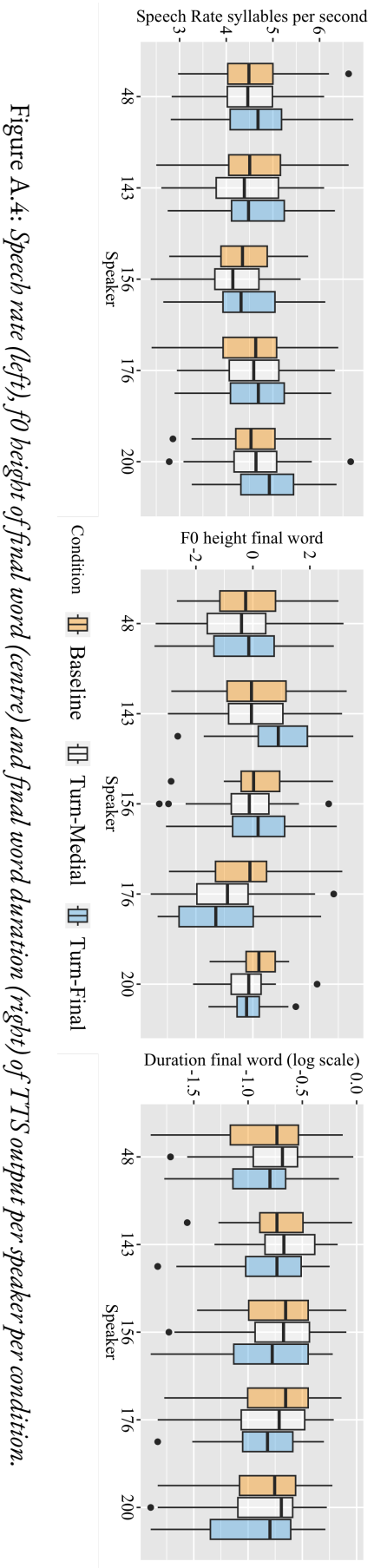


Figure A.4: *Speech rate (left), f0 height of final word (centre) and final word duration (right) of TTS output per speaker per condition.*

## A.6 Chapter 7: Textual Material for Listening Tests

Table A.6: *Textual Material used in Listening Tests for Chapter 7 (turn-ambiguous)*

| **Stimulus** | **Text** |
|---|---|
| 0 | a couple times |
| 1 | the troubleshooting one |
| 2 | i was like someone because i kept just looking i kept checking and they gave me one i think the limit was maybe one per customer |
| 3 | okay so you've got significantly better weather than me |
| 4 | but they're they're they're old boys |
| 5 | sending notes to old college friends through words with friends |
| 6 | norway and stuff like that just for the nature |
| 7 | um you know i just point to people like that you know who get so much out of it who get a community from it |
| 8 | but that's really cool that you're at least still working |
| 9 | and nothing nothing happened |
| 10 | dishes of mac and cheese that's going to be too much like no one's gonna finish that so i looked it up and you can actually make um mac and cheese |
| 11 | like he was on it for maybe two minutes he was having technical difficulties because i could not see him |
| 12 | yeah it's been crazy |
| 13 | like polar opposites of like the country as well i got chicago and they got houston once for snow one for heat |
| 14 | they would travel on trains sometimes but they weren't there was no like mass media really aside from i guess like yeah newspapers already existed |
| 15 | and then it just it just changed |
| 16 | i went for me and my dad went to the best western |
| 17 | south jersey is a lot slower |
| 18 | and they started dating |
| 19 | i'm good in you |
| 20 | of you know the finished product |
| 21 | no no it's gonna just keep getting worse and worse you know we go through the whole thing why it's going to happen but anyway |
| 22 | you're trying to connect z to a |
| 23 | and like we're the only people that we all hang out with |
| 24 | with a k |
| 25 | so i don't know what happened |
| 26 | yeah i love life theaters |
| 27 | they do work to a certain extent like they still work so if i glance at my watch and i'm not thinking super hard about it |
| 28 | his school went online and he's still been online and now he's planning to go away to college |
| 29 | like there's not a whole lot in between |
| 30 | now it's a black screen for you |

| | |
|---|---|
| 31 | you are not always like have to be on campus |
| 32 | yeah i mean i think it's pretty clear that here on the east coast like we missed it we were late |
| 33 | i don't see why they wouldn't want to |
| 34 | um the computer isn't the same |
| 35 | you make me a drink |
| 36 | i did that that's not really what i wanted to do i wanted to do the experience |
| 37 | heard smile dog |
| 38 | i thought i was going to be coming in here like awkward but i mean like i can make the conversation like once we start talking you know |
| 39 | um technically we're back to normal but not really |
| 40 | yeah the world is just crazy right now |
| 41 | four day complete lockdown |
| 42 | i keep people out |
| 43 | it's tough though because for me it's like really bring my motivation down because it's just like no change of scenery |
| 44 | and i don't know about you but i have an suv so like i'll do a little fish tailing sometimes |
| 45 | you would hope |
| 46 | the guy who started he purposely designed it to that |
| 47 | so so you usually end up getting the really critical cases i've heard |
| 48 | i actually just graduated to me |
| 49 | like the tree |

# Bibliography


Adigwe, A., Wallbridge, S., and King, S. (2024). What do people hear? Listeners' Perception of Conversational Speech. In *Proc. Interspeech 2024*, pages 1210–1214. ISCA.

Adigwe, A. O. and Klabbers, E. (2022). Strategies for developing a Conversational Speech Dataset for Text-To-Speech Synthesis. In *Proc. Interspeech 2022*, pages 2318–2322. ISCA.

Aijmer, K. (2015). Well in an English-Swedish and English-French Contrastive Perspective. In Beeching, K. and Woodfield, H., editors, *Researching Sociopragmatic Variability: Perspectives from Variational, Interlanguage and Contrastive Pragmatics*, pages 201–229. Palgrave Macmillan UK, London.

Aijmer, K. and Simon-Vandenbergen, A.-M. (2003). The discourse particle well and its equivalents in Swedish and Dutch. *Linguistics*, 41(6):1123–1161.

An, X., Wang, Y., Yang, S., Ma, Z., and Xie, L. (2019). Learning Hierarchical Representations for Expressive Speaking Style in End-to-End Speech Synthesis. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 184–191.

Andersson, S. (2013). *Synthesis and Evaluation of Conversational Characteristics in Speech Synthesis*. PhD thesis, University of Edinburgh, Edinburgh.

Andersson, S., Georgila, K., Traum, D., Aylett, M., and Clark, R. A. J. (2010). Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech for Unit Selection. In *Proc. Speech Prosody 2010*.

Andersson, S., Yamagishi, J., and Clark, R. A. J. (2012). Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication*, 54(2):175–188.

Aubin, A., Cervone, A., Watts, O., and King, S. (2019). Improving Speech Synthesis with Discourse Relations. In *Proc. Interspeech 2019*, pages 4470–4474. ISCA.

Aviad, A., Kaland, C., Ellison, T. M., Cangemi, F., Winter, B., and Grice, M. (2024). Harvesting spontaneous speech data from digital reservoirs to study prosody. In *Proceedings of the 19th Conference on Laboratory Phonology (LabPhon 19)*.

Badlani, R., Łańcucki, A., Shih, K. J., Valle, R., Ping, W., and Catanzaro, B. (2022). One TTS Alignment to Rule Them All. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6092–6096.

Baljekar, P. and Black, A. W. (2016). Utterance Selection Techniques for TTS Systems Using Found Speech. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pages 184–189. ISCA.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Batliner, A., Kompe, R., Kiessling, A., Noeth, E., and Niemann, H. (1995). Can You Tell Apart Spontaneous and Read Speech if You Just Look at Prosody? In Ayuso, A. J. R. and Soler, J. M. L., editors, *Speech Recognition and Coding: New Advances and Trends*, volume 147, pages 101–104. Springer.

Beckman, M. E. (1997). A Typology of Spontaneous Speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pages 7–26. Springer US, New York, NY.

Beckman, M. E., Hirschberg, J. B., and Shattuck-Hufnagel, S. (2005). Chapter 2: The Original ToBI System and the Evolution of the ToBI Framework. In Jun, S.-A., editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*, pages 9–54. Oxford Academic.

Ben-David, A. and Shechtman, S. (2021). Acquiring conversational speaking style from multi-speaker spontaneous dialog corpus for prosody-controllable sequence-to-sequence speech synthesis. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 66–71.

Betz, S., Carlmeyer, B., Wagner, P., and Wrede, B. (2018). Interactive Hesitation Synthesis: Modelling and Evaluation. *Multimodal Technologies and Interaction*, 2(1):9.

Bishop, J. (2012). Information structural expectations in the perception of prosodic prominence. In *Information structural expectations in the perception of prosodic prominence*, pages 239–270. De Gruyter Mouton.

Blakemore, D. (2002). Introduction. In *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge University Press, Cambridge.

Boersma, P. and Weenink, D. (2021). Praat: doing phonetics by computer (Version 6.1.38).

Boersma, P. and Weenink, D. (2024). Praat: doing phonetics by computer (Version 6.4.05).

Bolden, G. B. (2006). Little Words That Matter: Discourse Markers "So" and "Oh" and the Doing of Other-Attentiveness in Social Interaction. *Journal of Communication*, 56(4):661–688.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020). Pyannote.Audio: Neural Building Blocks for Speaker Diarization. In *Proc. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128.

Brinton, L. J. (1996). Conceptual Background. In *Pragmatic Markers in English: Grammaticalization and Discourse Functions*. De Gruyter Mouton, Berlin, New York.

Brown, M., Salverda, A. P., Gunlogson, C., and Tanenhaus, M. K. (2015). Interpreting prosodic cues in discourse context. *Language, cognition and neuroscience*, 30(1-2):149–166.

Brusco, P., Pérez, J., and Gravano, A. (2017). Cross-Linguistic Study of the Production of Turn-Taking Cues in American English and Argentine Spanish. In *Proc. Interspeech 2017*, pages 2351–2355.

Brusco, P., Vidal, J., Beňuš, , and Gravano, A. (2020). A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. *Speech Communication*, 125:24–40.

Bögels, S. and Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Bögels, S. and Torreira, F. (2021). Turn-end Estimation in Conversational Turn-taking: The Roles of Context and Prosody. *Discourse Processes*, 58(10):903–924.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Calhoun, S. and Schweitzer, A. (2012). Can intonation contours be lexicalised? Implications for discourse meanings. In Elordieta, G. and Prieto, P., editors, *Prosody and Meaning*, pages 271–328. De Gruyter.

Calhoun, S., Wollum, E., and Kruse Va'ai, E. (2021). Prosodic Prominence and Focus: Expectation Affects Interpretation in Samoan and English. *Language and Speech*, 64(2):346–380.

Camp, J., Kenter, T., Finkelstein, L., and Clark, R. (2023). MOS vs. AB: Evaluating Text-to-Speech Systems Reliably Using Clustered Standard Errors. In *Proc. Interspeech 2023*, pages 1090–1094. ISCA.

Campbell, N. (2005). Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language. *Language Resources and Evaluation*, 39(1):109–118.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1171–1178.

Campbell, W. N. (1997). Synthesizing Spontaneous Speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pages 165–186. Springer US, New York, NY.

Canavan, A., Graff, D., and Zipperlen, G. (1997). CALLHOME American English Speech. ISBN: 9781585631117 Publisher: Linguistic Data Consortium.

Cangemi, F. and Baumann, S. (2020). Integrating phonetics and phonology in the study of linguistic prominence. *Journal of Phonetics*, 81:100993.

Cangemi, F., Grice, M., Jeon, H.-S., and Setter, J. (2023). Contrast or Context, That is the Question. In *Proceedings of the 20th International Congress of Phonetic Sciences*, pages 1360–1364.

Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., and Weber, J. (2024). XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Proc. Interspeech 2024*, pages 4978–4982. ISCA.

Chen, L.-W., Watanabe, S., and Rudnicky, A. (2023). A vector quantized approach for text to speech synthesis on real-world spontaneous speech. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, volume 37, pages 12644–12652.

Chermaz, C. and King, S. (2020). A Sound Engineering Approach to Near End Listening Enhancement. In *Proc. Interspeech 2020*, pages 1356–1360. ISCA.

Chodroff, E. and Cole, J. S. (2019). Testing the Distinctiveness of Intonational Tunes: Evidence from Imitative Productions in American English. In *Proc. Interspeech 2019*, pages 1966–1970.

Christensen, R. H. B. (2022). Regression Models for Ordinal Data: Introducing R-package ordinal. R package version 2022.11-16.

Cieri, C., Graff, D., Kimball, O., Miller, D., and Walker, K. (2004). Fisher English Training Speech Part 1 Speech LDC2004S13.

Clark, H. H. (1996a). Conversation. In *Using Language*, Using Linguistic Books, pages 318–352. Cambridge University Press, Cambridge.

Clark, H. H. (1996b). Language use. In *Using Language*, 'Using' Linguistic Books, pages 3–26. Cambridge University Press, Cambridge.

Clark, R., Silen, H., Kenter, T., and Leith, R. (2019). Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 99–104. ISCA.

Clifton, A., Reddy, S., Yu, Y., Pappu, A., Rezapour, R., Bonab, H., Eskevich, M., Jones, G., Karlgren, J., Carterette, B., and Jones, R. (2020). 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917. International Committee on Computational Linguistics.

Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, 30(1-2):1–31.

Cole, J., Mo, Y., and Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2):425–452.

Cole, J., Steffman, J., and Awwad, A. (2024). Functional modeling of F0 variation across speakers and between phonological categories: Rising pitch accents in American English. In *Speech Prosody 2024*, pages 1020–1024. ISCA.

Cong, J., Yang, S., Hu, N., Li, G., Xie, L., and Su, D. (2021). Controllable Context-Aware Conversational Speech Synthesis. In *Proc. Interspeech 2021*, pages 4658–4662. ISCA.

Cooper, E., Huang, W.-C., Tsao, Y., Wang, H.-M., Toda, T., and Yamagishi, J. (2024). A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, 45(4):161–183.

Cooper, E., Wang, X., Chang, A., Levitan, Y., and Hirschberg, J. (2017). Utterance Selection for Optimizing Intelligibility of TTS Voices Trained on ASR Data. In *Proc. Interspeech 2017*, pages 3971–3975. ISCA.

Corkey, N., O'Mahony, J., and King, S. (2023). Intonation Control for Neural Text-to-Speech Synthesis with Polynomial Models of F0. In *Proc. Interspeech 2023*, pages 2014–2015. ISCA.

Couper-Kuhlen, E. (2005). Intonation and Discourse: Current Views from Within. In *The Handbook of Discourse Analysis*, pages 11–34. John Wiley & Sons, Ltd.

Couper-Kuhlen, E. and Selting, M. (1996). Towards an interactional perspective on prosody and a prosodic perspective on interaction. In Couper-Kuhlen, E. and Selting, M., editors, *Prosody in Conversation: Interactional Studies*, Studies in Interactional Sociolinguistics, pages 11–56. Cambridge University Press, Cambridge.

Cruttenden, A. (1984). The Relevance of Intonational Misfits. In Gibbon, y. and Richter, H., editors, *Intonation, Accent and Rhythm: Studies in Discourse Phonology*, pages 67–76. De Gruyter.

Cruttenden, A. (1997a). *Intonation*. Cambridge University Press, second edition.

Cruttenden, A. (1997b). Preliminaries. In *Intonation*, Cambridge Textbooks in Linguistics, pages 1–12. Cambridge University Press, Cambridge, second edition.

Culpeper, J. (2011). Chapter 2 "It's not what you said, it's how you said it!" Prosody and impoliteness. In *Discursive Approaches to Politeness, edited by Linguistic Politeness Research Group*, pages 57–84. De Gruyter Mouton.

Cutler, A. and Pearson, M. (1985). On The Analysis of Prosodic Turn-Taking Cues. In Johns-Lewis, C., editor, *Intonation in Discourse*, pages 139–156. Routledge.

Dall, R. (2017). *Statistical Parametric Speech Synthesis Using Conversational Data and Phenomena*. PhD thesis, University of Edinburgh.

Dall, R., Brognaux, S., Richmond, K., Valentini-Botinhao, C., Henter, G. E., Hirschberg, J., Yamagishi, J., and King, S. (2016a). Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis. In *Proc. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5155–5159. IEEE.

Dall, R., Tomalin, M., and Wester, M. (2016b). Synthesising Filled Pauses: Representation and Datamixing. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pages 7–13. ISCA.

Dall, R., Tomalin, M., Wester, M., Byrne, W., and King, S. (2014a). Investigating automatic & human filled pause insertion for speech synthesis. In *Proc. Interspeech 2014*, pages 51–55. ISCA.

Dall, R., Yamagishi, J., and King, S. (2014b). Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation. In *Proc. Speech Prosody 2014*, pages 1012–1016. ISCA.

de Jong, K., Beckman, M. E., and Edwards, J. (1993). The Interplay Between Prosodic Structure and Coarticulation. *Language and Speech*, 36(2-3):197–212.

de Looze, C. and Rauzy, S. (2009). Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register variations and Pause Duration. In *Proc. Interspeech 2009*, pages 2919–2922. ISCA.

Degand, L. and Simon, A. C. (2009). On identifying basic discourse units in speech: theoretical and empirical issues. *Discours*, (4).

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834. ISCA.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *lournal of Personality and Social Psychology*, 23(2):283–292.

Edlund, J., Gustafson, J., Heldner, M., and Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication*, 50(8):630–645.

Edlund, J. and Heldner, M. (2005). Exploring Prosody in Interaction Control. *Phonetica*, 62(2-4):215–226. De Gruyter Mouton.

Ekstedt, E. and Skantze, G. (2022a). How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In *Proc. SIGdial 2022 Conference*, pages 541–551.

Ekstedt, E. and Skantze, G. (2022b). Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*, pages 5190–5194. ISCA.

Ekstedt, E., Wang, S., Székely, , Gustafson, J., and Skantze, G. (2023). Automatic Evaluation of Turn-taking Cues in Conversational Speech Synthesis. In *Proc. Interspeech 2023*, pages 5481–5485. ISCA.

Ernestus, M. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39(3):253–260.

Evanini, K. and Lai, C. (2010). The importance of optimal parameter setting for pitch extraction. *The Journal of the Acoustical Society of America*, 128:2291–2291.

Farrús, M., Lai, C., and Moore, J. D. (2016). Paragraph-based prosodic cues for speech synthesis applications. In *Proc. Speech Prosody 2016*, pages 1143–1147. ISCA.

Figueroa, C., Adigwe, A., Ochs, M., and Skantze, G. (2022). Annotation of Communicative Functions of Short Feedback Tokens in Switchboard. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1849–1859.

Freeman, V., Levow, G.-A., Wright, R., and Ostendorf, M. (2015a). Investigating the role of 'yeah' in stance-dense conversation. In *Proc. Interspeech 2015*, pages 3076–3080. ISCA.

Freeman, V., Wright, R., and Levow, G.-A. (2015b). The prosody of negative yeah. *LSA Annual Meeting Extended Abstracts*, 6:1–5.

Garrod, S. and Pickering, M. J. (2009). Joint Action, Interactive Alignment, and Dialog. *Topics in Cognitive Science*, 1(2):292–304.

Gillick, J., Deng, W., Ryokai, K., and Bamman, D. (2021). Robust Laughter Detection in Noisy Environments. In *Proc. Interspeech 2021*, pages 2481–2485. ISCA.

Godfrey, G. and Holliman, E. (1993). Switchboard-1 Release 2. Linguistic Data Consortium.

Goodwin, C. and Duranti, A. (1992). Rethinking context: an introduction. In *Rethinking context: Language as an interactive phenomenon*, pages 1–42. Cambridge University Press.

Govender, A., Valentini-Botinhao, C., and King, S. (2019). Measuring the contribution to cognitive load of each predicted vocoder speech parameter in DNN-based speech synthesis. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 121–126. ISCA.

Grabe, E., Kochanski, G., and Coleman, J. (2003). Quantitative modelling of intonational variation. In *Proceedings of SASRTLM 2003 (Speech Analysis and Recognition in Technology, Linguistics and Medicine)*.

Grabe, E., Kochanski, G., and Coleman, J. (2005). The intonation of native accent varieties in the British Isles: Potential for Miscommunication? *English Pronunciation Models: A Changing Scene*, pages 311–337.

Grabe, E., Kochanski, G., and Coleman, J. (2006). Empirical Validation of Hand-labelled Nuclear Accent Patterns. In *Proc. Speech Prosody 2006*, pages paper–020.

Grabe, E., Kochanski, G., and Coleman, J. (2007). Connecting Intonation Labels to Mathematical Descriptions of Fundamental Frequency. *Language and Speech*, 50(3):281–310.

Gravano, A., Brusco, P., and Beňuš, (2016). Who Do You Think Will Speak Next? Perception of Turn-Taking Cues in Slovak and Argentine Spanish. In *Proc. Interspeech 2016*, pages 1265–1269. ISCA.

Gravano, A. and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Gravano, A., Hirschberg, J., and Beňuš, (2011). Affirmative Cue Words in Task-Oriented Dialogue. *Computational Linguistics*, 38(1):1–39.

Gravano, A. and Vidal, C. A. J. (2014). A Study of Turn-Yielding Cues in Human-Computer Dialogue. In *Proc. 15th Argentine Symposium on Articial Intelligence (ASAI)*, pages 9–17.

Gumperz, J. J. (1992). Contextualizaion and understanding. In *Rethinking context: Language as an interactive phenomenon*, pages 229–252. Cambridge University Press, Great Britain.

Guo, H., Zhang, S., Soong, F., He, L., and Xie, L. (2021). Conversational End-to-End TTS for Voice Agents. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 403–409.

Gutierrez, E., Oplustil-Gallegos, P., and Lai, C. (2021). Location, Location: Enhancing the Evaluation of Text-to-Speech synthesis using the Rapid Prosody Transcription Paradigm. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 25–30. ISCA.

Hazan, V. and Baker, R. (2010). Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style? In *Proceedingd of DiSS-LPSS Joint Workshop*, pages 7–10, Tokyo, Japan.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Heldner, M., Wlodarczak, M., Beňuš, , and Gravano, A. (2019). Voice Quality as a Turn-Taking Cue. In *Proc. Interspeech 2019*, pages 4165–4169.

Helt, M. E. (1997). *Discourse marker and stance adverbial variation in spoken American English: A corpus-based analysis*. Ph.D., Northern Arizona University, United States – Arizona. ISBN: 9780591337709.

Heritage, J. (1984). Conversation Analysis. In Heritage, J., editor, *Garfinkel and Ethnomethodology*, page 233. Cambridge Polity Press.

Heritage, J. (2015). Well-prefaced turns in English conversation: A conversation analytic perspective. *Journal of Pragmatics*, 88:88–104.

Hirschberg, J. (1990). Using discourse context to guide pitch accent decisions in synthetic speech. In *Proc. First ESCA Workshop on Speech Synthesis (SSW 1)*, pages 181–184.

Hirschberg, J. and Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3).

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1):23–35.

Hjalmarsson, A. and Laskowski, K. (2011). Measuring final lengthening for speaker-change prediction. In *Proc. Interspeech 2011*, pages 2065–2068. ISCA.

Hodari, Z. (2022). *Synthesising prosody with insufficient context*. PhD thesis, University of Edinburgh.

Hodari, Z., Lai, C., and King, S. (2020). Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0. In *Proc. Speech Prosody 2020*, pages 965–969. ISCA.

Hodari, Z., Watts, O., and King, S. (2019). Using generative modelling to produce varied intonation for speech synthesis. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 239–244.

House, J. (2006). Constructing a context with intonation. *Journal of Pragmatics*, 38(10):1542–1558.

House, J. (2007). The role of prosody in constraining context selection: a procedural approach. *Cahiers de Linguistique Francaise*, 2:369–383.

Howcroft, D. M. and Rieser, V. (2021). What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939. Association for Computational Linguistics.

Howell, P. and Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication*, 10(2):163–169.

Hsia, C.-C., Wu, C.-H., and Wu, J.-Y. (2010). Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):1994–2003.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Hu, J. and Degand, L. (2023). The Conversational Discourse Unit: Identification and Its Role in Conversational Turn-taking Management. *Dialogue & Discourse*, 14(2):83–112.

Hu, Y., Liu, R., Gao, G., and Li, H. (2022). FCTalker: Fine and Coarse Grained Context Modeling for Expressive Conversational Speech Synthesis. arXiv:2210.15360 [cs, eess].

Im, S., Cole, J., and Baumann, S. (2023). Standing out in context: Prominence in the production and perception of public speech. *Laboratory Phonology*, 24(1).

Innes, B. (2010). "Well, That's Why I Asked the Question Sir": Well as a Discourse Marker in Court. *Language in Society*, 39(1):95–117. Cambridge University Press.

Ito, K. and Johnson, L. (2017). The LJ Speech Dataset.

Jadoul, Y., Thompson, B., and Boer, B. d. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Kakouros, S. and O'Mahony, J. (2023). What does BERT learn about prosody? In *Proc. 20th International Congress of Phonetic Sciences*, pages 1454–1458. Guarant International.

Kakouros, S., Šimko, J., Vainio, M., and Suni, A. (2023). Investigating the Utility of Surprisal from Large Language Models for Speech Synthesis Prosody: Speech Synthesis Workshop. In Hueber, T., Lolive, D., Obin, N., and Perrotin, O., editors, *Proc. 12th ISCA Speech Synthesis Workshop (SSW)*, pages 127–133. ISCA.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S., and Kavukcuoglu, K. (2018). Efficient Neural Audio Synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2410–2419. PMLR.

Kim, J., Kong, J., and Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *Proc. International Conference on Machine Learning (ICML)*.

Kirkland, A., Mehta, S., Lameris, H., Henter, G. E., Szekely, E., and Gustafson, J. (2023). Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *Proc. SSW 2023*, pages 41–47. ISCA.

Kirkland, A., Włodarczak, M., Gustafson, J., and Szekely, E. (2021). Perception of smiling voice in spontaneous speech synthesis. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 108–112. ISCA.

Klimkov, V., Nadolski, A., Moinet, A., Putrycz, B., Barra-Chicote, R., Merritt, T., and Drugman, T. (2017). Phrase Break Prediction for Long-Form Reading TTS: Exploiting Text Structure Information. In *Proc. Interspeech 2017*, pages 1064–1068. ISCA.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998). An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs. *Language and Speech*, 41(3-4):295–321.

Kominek, J. and Black, A. W. (2004). The CMU Arctic Speech Databases. In *Proc. 5th ISCA Speech Synthesis Workshop*, pages 223–224. ISCA.

Kong, J., Kim, J., and Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proc. NeurIPS*, volume 33, pages 17022–17033.

Kontogiorgos, D., Avramova, V., Alexanderson, S., Jonell, P., Oertel, C., Beskow, J., Skantze, G., and Gustafson, J. (2018). A Multimodal Corpus for Mutual Gaze and Joint Attention in Multiparty Situated Interaction. In *Proc. Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 119–127.

Koriyama, T., Nose, T., and Kobayashi, T. (2010). Conversational spontaneous speech synthesis using average voice model. In *Proc. Interspeech 2010*, pages 853–856. ISCA.

Kruijff-Korbayová, I., Ericsson, S., Rodríguez, K. J., and Karagjosova, E. (2003). Producing contextually appropriate intonation in an information-state based dialogue system. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - EACL '03*, volume 1, page 227. Association for Computational Linguistics.

Lai, C. (2014). Interpreting Final Rises: Task and Role Factors. In *Proc. Speech Prosody 2014*, pages 520–524. ISCA.

Lam, P. W. Y. (2009). What a Difference the Prosody Makes: The Role of Prosody in the Study of Discourse Particles. In Barth-Weingarten, D., Dehé, N., and Wichmann, A., editors, *Where Prosody Meets Pragmatics*, pages 107–126. Brill.

Lameris, H., Kirkland, A., Gustafson, J., and Szekely, E. (2023). Situating Speech Synthesis: Investigating Contextual Factors in the Evaluation of Conversational TTS. In *Proc. 12th Speech Synthesis Workshop (SSW) 2023*.

Lameris, H., Mehta, S., Henter, G. E., Gustafson, J., and Székely, (2022). Prosody-controllable spontaneous TTS with neural HMMs. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Lameris, H., Szekely, E., and Gustafson, J. (2024). The Role of Creaky Voice in Turn Taking and the Perception of Speaker Stance: Experiments Using Controllable TTS. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16058–16065.

Latorre, J., Lachowicz, J., Lorenzo-Trueba, J., Merritt, T., Drugman, T., Ronanki, S., and Klimkov, V. (2019). Effect of Data Reduction on Sequence-to-sequence Neural TTS. In *Proc. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7075–7079. IEEE.

Lavechin, M., Métais, M., Titeux, H., Boissonnet, A., Copet, J., Rivière, M., Bergelson, E., Cristia, A., Dupoux, E., and Bredin, H. (2023). Brouhaha: Multi-Task Training for Voice Activity Detection, Speech-to-Noise Ratio, and C50 Room Acoustics Estimation. In *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Lee, K., Park, K., and Kim, D. (2023). DailyTalk: Spoken Dialogue Dataset for Conversational Text-to-Speech. In *Proc. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Lee, L., Jouvet, D., Bartkova, K., Keromnes, Y., and Dargnat, M. (2020). Correlation Between Prosody and Pragmatics: Case Study of Discourse Markers in French and English. In *Proc. Interspeech 2020*, pages 1878–1882. ISCA.

Lee, L., Tseng, C., and Hsieh, C. (1993). Improved tone concatenation rules in a formant-based Chinese text-to-speech system. *IEEE Transactions on Speech and Audio Processing*, 1(3):287–294. Conference Name: IEEE Transactions on Speech and Audio Processing.

Li, J., Meng, Y., Li, C., Wu, Z., Meng, H., Weng, C., and Su, D. (2022). Enhancing Speaking Styles in Conversational Text-to-Speech Synthesis with Graph-Based Multi-Modal Context Modeling. In *Proc. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7917–7921, Singapore.

Li, Z., Zhang, Y., Nie, M., Yan, M., He, M., Zhang, R., and Gong, C. (2021). Improving Prosody for Unseen Texts in Speech Synthesis by Utilizing Linguistic Information and Noisy Data. arXiv:2111.07549 [cs, eess].

Loizou, P. C. (2011). Speech Quality Assessment. In Kacprzyk, J., Lin, W., Tao, D., Kacprzyk, J., Li, Z., Izquierdo, E., and Wang, H., editors, *Multimedia Analysis, Processing and Communications. Studies in Computational Intelligence*, volume 346, pages 623–654. Springer, Berlin, Heidelberg.

Luong, H.-T., Wang, X., Yamagishi, J., and Nishizawa, N. (2019). Training Multi-Speaker Neural Text-to-Speech Systems Using Speaker-Imbalanced Speech Corpora. In *Proc. Interspeech 2019*, pages 1303–1307. ISCA.

Lyth, D. and King, S. (2024). Natural language guidance of high-fidelity text-to-speech with synthetic annotations. arXiv:2402.01912 [cs, eess].

Malisz, Z., Henter, G. E., Botinhao, C. V., Watts, O., Beskow, J., and Gustafson, J. (2019). Modern speech synthesis for phonetic sciences: a discussion and an evaluation. In *Proc. 19th International Congress of Phonetic Sciences ICPhS 2019*, pages 487–491.

Mary, L., Babu K. K, A., Joseph, A., and George, G. M. (2013). Evaluation of mimicked speech using prosodic features. In *Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7189–7193. IEEE.

Mauch, M. and Dixon, S. (2014). PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663. IEEE.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., Niekirk, B. v., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., Halvachs, E., Thomé, C., Robert-Stöter, F., Bittner, R., Wei, Z., Weiss, A., Battenberg, E., Choi, K., Yamamoto, R., Carr, C. J., Metsai, A., Sullivan, S., Friesch, P., Krishnakumar, A., Hidaka, S., Kowalik, S., Keller, F., Mazur, D., Chabot-Leclerc, A., Hawthorne, C., Ramaprasad, C., Keum, M., Gomez, J., Monroe, W., Morozov, V. A., Eliasi, K., nullmightybofo, Biberstein, P., Sergin, N. D., Hennequin, R., Naktinis, R., beantowel, Kim, T., Åsen, J. P., Lim, J., Malins, A., Hereñú, D., Struijk, S. v. d., Nickel, L., Wu, J., Wang, Z., Gates, T., Vollrath, M., Sarroff, A., Xiao-Ming, Porter, A., Kranzler, S., Voodoohop, Gangi, M. D., Jinoz, H., Guerrero, C., Mazhar, A., toddrme2178, Baratz, Z., Kostin, A., Zhuang, X., Lo, C. T., Campr, P., Semeniuc, E., Biswal, M., Moura, S., Brossier, P., Lee, H., and Pimenta, W. (2023). librosa/librosa: 0.10.1.

Michilsen, V. E. (2019). *On the relation between the prosody and discourse functions of well*. Bachelor Thesis, Utrecht University.

Mitsui, K., Zhao, T., Sawada, K., Hono, Y., Nankaku, Y., and Tokuda, K. (2022). End-to-End Text-to-Speech Based on Latent Representation of Speaking Styles Using Spontaneous Dialogue. In *Proc. Interspeech 2022*, pages 2328–2332. ISCA.

Mo, Y., Cole, J., and Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: evidence from formant structure. In *Interspeech 2009*, pages 2535–2538. ISCA.

Möbius, B. (2003). Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis. *International Journal of Speech Technology*, 6(1):57–71.

Möhler, G. and Conkie, A. (1998). Parametric modeling of intonation using vector quantization. In *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis (SSW 3)*, pages 311–316.

Nishimura, Y., Saito, Y., Takamichi, S., Tachibana, K., and Saruwatari, H. (2022). Acoustic Modeling for End-to-End Empathetic Dialogue Speech Synthesis Using Linguistic and Prosodic Contexts of Dialogue History. In *Proc. Interspeech 2022*, pages 3373–3377.

Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2):204–238.

Ogden, R. (2006). Phonetics and social action in agreements and disagreements. *Journal of Pragmatics*, 38(10):1752–1775.

Ogden, R. (2012). Making Sense of Outliers. *Phonetica*, 69(1-2):48–67. De Gruyter Mouton.

Ogden, R. A. (2007). Details and Contexts. In *Proc. ICPhS XVI*, pages 215–218.

Oplustil-Gallegos, P. and King, S. (2020). Using previous acoustic context to improve Text-to-Speech synthesis. arXiv:2012.03763 [cs].

Oplustil-Gallegos, P., O'Mahony, J., and King, S. (2021). Comparing acoustic and textual representations of previous linguistic context for improving Text-to-Speech. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 205–210. ISCA.

Park, Kyubyong & Kim, J. (2019). g2pE. Publication Title: GitHub repository.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., and Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pierrehumbert, J. (1981). Synthesizing intonation. *The Journal of the Acoustical Society of America*, 70(4):985–995.

Piits, L., Pajupuu, H., Sahkai, H., Altrov, R., Ermus, L., Tamuri, K., Hein, I., Mihkla, M., Kiissel, I., Männisalu, E., Suluste, K., and Pajupuu, J. (2022). Audiobook Dialogues as Training Data for Conversational Style Synthetic Voices. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

Popescu-Belis, A. and Zufferey, S. (2011). Automatic identification of discourse markers in dialogues: An in-depth study of like and well. *Computer Speech & Language*, 25(3):499–518.

Prevost, S. and Steedman, M. (1994). Specifying intonation from context for speech synthesis. *Speech Communication*, 15(1-2):139–153.

Purse, R. and Krivokapić, J. (2023). The Kinematic Properties of Prosodic Boundaries in Conversational Turn-taking. In *Proc. 20th International Congress of Phonetic Sciences*, pages 1653–1657.

Raitio, T., Li, J., and Seshadri, S. (2022a). Hierarchical Prosody Modeling and Control in Non-Autoregressive Parallel Neural TTS. *Proc. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7587–7591.

Raitio, T., Petkov, P., Li, J., Shifas, M., Davis, A., and Stylianou, Y. (2022b). Vocal effort modeling in neural TTS for improving the intelligibility of synthetic speech in noise. In *Proc. Interspeech 2022*, pages 1936–1940. ISCA.

Raitio, T., Rasipuram, R., and Castellani, D. (2020). Controllable neural text-to-speech synthesis using intuitive prosodic features. In *Proc. Interspeech 2020*, pages 4432–4436. ISCA.

Rakov, R. (2019). *Analyzing Prosody with Legendre Polynomial Coefficients*. PhD thesis, The City University of New York.

Rakov, R. and Rosenberg, A. (2017). Investigating native and non-native English classification and transfer effects using Legendre polynomial coefficient clustering. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 637–643.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., De Mori, R., and Bengio, Y. (2021). SpeechBrain: A General-Purpose Speech Toolkit. arXiv:2106.04624 [cs, eess].

Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A., and Marin, S. (2023). The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13).

Reichel, U. (2011). The CoPaSul intonation model. In *Sprachsignalverarbeitung, Spracherkennung und Sprachsynthese II*, pages 341–348.

Reichel, U. D. (2007). Data-driven Extraction of Intonation Contour Classes. In *Proc. 6th ISCA Workshop on Speech Synthesis,*, pages 240–245.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2021). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2019). FastSpeech: Fast, Robust and Controllable Text to Speech. In *Proc. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019),*, Vancouver, British Columbia, Canada.

Roettger, T. B. and Rimland, K. (2020). Listeners' adaptation to unreliable intonation is speaker-sensitive. *Cognition*, 204:104372.

Romero-Trillo, J. (2018). Prosodic modeling and position analysis of pragmatic markers in English conversation. *Corpus Linguistics and Linguistic Theory*, 14(1):169–195. De Gruyter Mouton.

Rosenberg, A. and Ramabhadran, B. (2017). Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores. In *Proc. Interspeech 2017*, pages 3976–3980. ISCA.

Ross, A., Corley, M., and Lai, C. (2024). Is there an uncanny valley for speech? Investigating listeners' evaluations of realistic TTS voices. In *Proc. Speech Prosody 2024*, pages 1115–1119. ISCA.

Ruiter, J.-P. D., Mitterer, H., and Enfield, N. J. (2006). Projecting the End of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Language*, 82(3):515–535.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735. Publisher: Linguistic Society of America.

Saeki, T., Wang, G., Morioka, N., Elias, I., Kastner, K., Rosenberg, A., Ramabhadran, B., Zen, H., Beaufays, F., and Shemtov, H. (2024). Extending Multilingual Speech Synthesis to 100+ Languages without Transcribed Data. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Sakita, T. I. (2017). Stance management in oral narrative: The role of discourse marker *well* and resonance. *Functions of Language*, 24(1):65–93.

Schegloff, E. A. (1980). What type of interaction is it to be. In *Proceedings of the 18th annual meeting on Association for Computational Linguistics*, pages 81–82, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Schegloff, E. A. (1999). Discourse, Pragmatics, Conversation, Analysis. *Discourse Studies*, 1(4):405–435. SAGE Publications.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., and Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9):1141–1152.

Schiffrin, D. (1990). Conversation Analysis. *Annual Review of Applied Linguistics*, 11:3–16.

Schuetze-Coburn, S. (1992). Prosodic Phrase as a Prototype. In *Proc. IRCS Workshop on Prosody in Natural Speech,*, pages 171–180.

Schweitzer, A., Möbius, B., Möhler, G., and Dogil, G. (2022). The PaIntE Model of Intonation. In *Prosodic Theory and Practice*, pages 351–375. The MIT Press.

Seebauer, F., Kuhlmann, M., Haeb-Umbach, R., and Wagner, P. (2023). Re-examining the quality dimensions of synthetic speech. In *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*, pages 34–40.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *Proc. ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Skantze, G. (2007). *Error handling in spoken dialogue systems: managing uncertainty, grounding and miscommunication*. PhD thesis, KTH Computer Science and Communication, Stockholm.

Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67:101178.

Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., and Saurous, R. A. (2018). Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. In *International Conference on Machine Learning*, pages 4700–4709.

Stan, A. and O'Mahony, J. (2023). An analysis on the effects of speaker embedding choice in non auto-regressive TTS. In *12th ISCA Speech Synthesis Workshop (SSW2023)*, pages 134–138. ISCA.

Stephens, J. and Beattie, G. (1986). On Judging the Ends of Speaker Turns in Conversation. *Journal of Language and Social Psychology*, 5(2):119–134. SAGE Publications.

Stephenson, B., Besacier, L., Girin, L., and Hueber, T. (2022). BERT, can HE predict contrastive focus? Predicting and controlling prominence in neural TTS using a language model. In *Proc. Interspeech 2022*, pages 3383–3387. ISCA.

Suni, A., Kakouros, S., Vainio, M., and Šimko, J. (2020). Prosodic Prominence and Boundaries in Sequence-to-Sequence Speech Synthesis. In *Proc. Speech Prosody 2020*, pages 940–944.

Suni, A., Šimko, J., Aalto, D., and Vainio, M. (2017). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Svatošová, M. and Volín, J. (2023). Description of F0 contours with Legendre polynomials. *AUC PHILOLOGICA*, 2022(1):97–113.

Swerts, M. and Hirschberg, J. (1998). Prosody and Conversation: An Introduction. *Language and Speech*, 41(3-4):229–233.

Syrdal, A. K. and McGory, J. (2000). Inter-transcriber reliability of toBI prosodic labeling. In *6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages 235–238. ISCA.

Szczepek Reed, B. (2010). Units of interaction:'Intonation phrases' or 'turn constructional phrases'? *Actes d'IDP (Interface Discours & Prosodie) 2009*.

Szekely, E., Henter, G. E., and Gustafson, J. (2019). Casting to Corpus: Segmenting and Selecting Spontaneous Dialogue for Tts with a CNN-LSTM Speaker-dependent Breath Detector. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6925–6929. IEEE.

Székely, , Eje Henter, G., Beskow, J., and Gustafson, J. (2019a). How to train your fillers: uh and um in spontaneous speech synthesis. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 245–250. ISCA.

Székely, , Henter, G. E., Beskow, J., and Gustafson, J. (2019b). Spontaneous Conversational Speech Synthesis from Found Data. In *Interspeech 2019*, pages 4435–4439. ISCA.

Tachibana, H., Uenoyama, K., and Aihara, S. (2018). Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. In *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788. IEEE Press.

Talman, A., Suni, A., Celikkanat, H., Kakouros, S., Tiedemann, J., and Vainio, M. (2019). Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations. In Hartmann, M. and Plank, B., editors, *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 281–290.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., Zhao, S., Qin, T., Soong, F., and Liu, T.-Y. (2024). NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4234–4245.

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *The Journal of the Acoustical Society of America*, 107(3):1697–1714.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press, first edition.

Tucker, B. V. and Mukai, Y. (2023). *Spontaneous Speech*. Cambridge University Press, first edition.

Tyagi, S., Nicolis, M., Rohnke, J., Drugman, T., and Lorenzo-Trueba, J. (2020). Dynamic Prosody Generation for Speech Synthesis Using Linguistics-Driven Acoustic Embedding Selection. In *Proc. Interspeech 2020*, pages 4407–4411. ISCA.

Van De Vreken, E., Richmond, K., and Lai, C. (2022). Voice Puppetry with FastPitch. In *Proc. Interspeech 2022*, pages 5219–5220. ISCA.

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, , Tånnander, C., and Voße, J. (2019). Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program. In *10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 105–110. ISCA.

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., Mancebo, D. E., Fivela, B. G., Lacheret, A., Moniz, H., Chasaide, A. N., Niebuhr, O., Rousier-Vercruyssen, L., Simon, A.-C., Šimko, J., Tesser, F., and Vainio, M. (2015a). Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence. In *Proceedings of the 18th International Congress of Phonetic Sciences. The Scottish Consortium for ICPhS 2015*.

Wagner, P., Trouvain, J., and Zimmerer, F. (2015b). In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48:1–12.

Wallbridge, S., Bell, P., and Lai, C. (2021). It's Not What You Said, it's How You Said it: Discriminative Perception of Speech as a Multichannel Communication System. In *Proc. Interspeech 2021*, pages 2386–2390. ISCA.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech 2017*, pages 4006–4010. ISCA.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., and Saurous, R. A. (2018). Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5180–5189. PMLR.

Watts, O., Wu, Z., and King, S. (2015). Sentence-level control vectors for deep neural network speech synthesis. In *Proc. Interspeech 2015*, pages 2217–2221. ISCA.

Wells, D., Richmond, K., and Lamb, W. (2023). A Low-Resource Pipeline for Text-to-Speech from Found Data With Application to Scottish Gaelic. In *Proc. Interspeech 2023*, pages 4324–4328. ISCA.

Wester, M., Aylett, M., Tomalin, M., and Dall, R. (2015). Artificial personality and disfluency. In *Proc. Interspeech 2015*, pages 3365–3369. ISCA.

Wilson, D. and Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, 38(10):1559–1579.

Wlodarczak, M. and Heldner, M. (2022). Contribution of voice quality to prediction of turn-taking events. In *Proc. Speech Prosody 2022*, pages 485–489.

Wu, C.-H., Hsia, C.-C., Lee, C.-H., and Lin, M.-C. (2010). Hierarchical Prosody Conversion Using Regression-Based Clustering for Emotional Speech Synthesis. *Proc. IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1394–1405.

Wu, Z., Watts, O., and King, S. (2016). Merlin: An Open Source Neural Network Speech Synthesis System. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pages 202–207.

Włodarczak, M. and Wagner, P. (2013). Effects of talk-spurt silence boundary thresholds on distribution of gaps and overlaps. In *Proc. Interspeech 2013*, pages 1434–1437. ISCA.

Xin, D., Adavanne, S., Ang, F., Kulkarni, A., Takamichi, S., and Saruwatari, H. (2023). Improving Speech Prosody of Audiobook Text-To-Speech Synthesis with Acoustic and Textual Contexts. In *Proc. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xu, G., Song, W., Zhang, Z., Zhang, C., He, X., and Zhou, B. (2021). Improving Prosody Modelling with Cross-Utterance Bert Embeddings for End-to-End Speech Synthesis. In *Proc. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6079–6083.

Xue, L., Soong, F. K., Zhang, S., and Xie, L. (2022). ParaTTS: Learning Linguistic and Prosodic Cross-Sentence Information in Paragraph-Based TTS. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 30:2854–2864.

Yamazaki, Y., Chiba, Y., Nose, T., and Ito, A. (2020). Construction and Analysis of a Multimodal Chat-talk Corpus for Dialog Systems Considering Interpersonal Closeness. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 443–448. European Language Resources Association.

Yamazaki, Y., Chiba, Y., Nose, T., and Ito, A. (2021). Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems. In *Proc. Interspeech 2021*, pages 246–250. ISCA.

Yang, L.-c. (2002). Interpreting meaning from context: modeling the prosody of discourse markers in speech. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 1193–1196. ISCA.

Yuan, J., Brenier, J. M., and Jurafsky, D. (2005). Pitch accent prediction: effects of genre and speaker. In *Proc. Interspeech 2005*, pages 1409–1412. ISCA.

Zandie, R., Mahoor, M. H., Madsen, J., and Emamian, E. S. (2021). RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis. In *Proc. Interspeech 2021*, pages 2751–2755. ISCA.

Zaïdi, J., Seuté, H., van Niekerk, B., and Carbonneau, M.-A. (2022). Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis. In *Proc. Interspeech 2022*, pages 4591–4595. ISCA.

Zellers, M. (2017). Prosodic Variation and Segmental Reduction and Their Roles in Cuing Turn Transition in Swedish. *Language and Speech*, 60(3):454–478. SAGE.

Zellers, M. and Ogden, R. (2014). Exploring Interactional Features with Prosodic Patterns. *Language and Speech*, 57(3):285–309.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (2019). LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech 2019*, pages 1526–1530. ISCA.

Łańcucki, A. (2021). Fastpitch: Parallel Text-to-Speech with Pitch Prediction. In *Proc. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592.

Šimko, J., Törö, T., Vainio, M., and Suni, A. (2023). Prosody under control: Controlling prosody in text-to-speech synthesis by adjustments in latent reference space. In *Proc. 20th International Congress of Phonetic Sciences*, pages 3086–3090.